ENHANCED IDENTIFICATION MODEL TO IMPROVE DATA LEAKAGE PREVENTION SYSTEMS IN OIL & GAS INDUSTRY

BY

ASHRAF ELTAHIR AHMED

A Thesis Presented to the

DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

December, 2017

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Ashraf Eltahir Mohamed Ahmed** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree **MASTER OF SCIENCE IN COMPUTER SCIENCE.**

Dr. Farag Azzedin
(Advisor)

Dr. Khalid A. Aljasser
Department Chairman

Dr. Lahouari Ghouti
(Member)

Dr. Salam A. Zummo
Dean of Graduate Studies

Dr. Sami Zhioua
(Member)

14/5/13
Date

Dedication

I would like to dedicate my thesis work to my beloved family. A special feeling of gratitude to my wife for her encouragements, care, love, and continues support throughout pursuing my master degree.

I would like also to thank Dr Farag Azzedin for his exceptional motivation, guidance, and continuous support. This thesis will not be accomplished without his presence. Finally, I would like to thank friends whom showed support and backed me up

ACKNOWLEDGMENTS

First of all, I would like to thank almighty Allah for giving me the opportunity to complete my master degree.

My sincere gratitude is expressed to my advisor Dr. Farag Azzedin for his guidance and support he provided throughout this research. I feel very fortunate to have had the opportunity to work under his supervision. Dr Azzedin devoted his time and energy to help me complete this research, he had been welcoming person, every walk to his office was pleasurable, I learned a lot with his accompany, I no longer see Dr Azzedin as a teacher only but as brother and a friend.

I would also like to thank the committee members, Dr. Lahouari Ghouti and Dr. Sami Zhioua and for dedicating time out of their busy schedules for this work and providing their feedback and comments. I also would like to thank other professors and graduate students who participated by comments and feedback.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DLP	:	Data Leakage Prevention
SVM	:	Support Vector Machine
IDS	:	Intrusion Detection Systems
IPS	:	Intrusion Prevention Systems
NIDSINIPS	:	Network based intrusion/prevention systems
HIDS/HIPS	:	Host based systems
SEIEM	:	Security Information Event Management
DCI	:	Deep Content Inspection
LM	:	Learning Method
NGFW	:	Next Generation Firewall
DST	:	Delineation Stem Test
PTA	:	Pressure Transient Analysis

ABSTRACT

Full Name : [Ashraf Eltahir Mohamed Ahmed]

Thesis Title : [ENHANCED IDENTIFICATION MODEL TO IMPROVE DATA LEAKAGE PREVENTION SYSTEMS IN OIL & GAS INDUSTRY]

Major Field : [ICS]

Date of Degree : [December 2017]

Data Leakage Prevention (DLP) is a technology that protects against potential data breach incidents throughout different stages in information lifecycle in timely manner. DLP technologies can be employed in organizations where access to sensitive information is required by a relatively large number of end-users. It can play an important role in preventing unauthorized distribution of such information and mitigate insider threat. It will monitor information interchange operations and apply defined policies against suspicious incidents. However, DLP technologies suffer some showstoppers: (a) difficulties in accurately describing information to be monitored, (b) extended implementation time, and (c) limitations in handling encrypted and/or graphical information.

DLP technologies can be improved by enhancing the monitoring activities so that less false positives and timely response are achieved. Improving DLP is a necessity for Oil & Gas industry because of the huge amount of sensitive information that is

regularly handled; otherwise, alerts raised by the DLP will be overlooked and lead to a false feeling of protection. In this research, we suggest a model to improve DLP in Oil & Gas companies by enhancing the identification of sensitive content and distributing the processing overhead over multiple stages. Our model will result in less number of false positive alerts, better use of the resources and less time to configure DLP systems.

ملخص الرسالة

الاسم الكامل: أشرف الطاهر محمد أحمد

عنوان الرسالة: نموذج محسن لتقنية منع تسرب البيانات

التخصص: علوم الحاسب و المعلومات

تاريخ الدرجة العلمية: ديسمبر 2017

منع تسرب البيانات هي تقنية حديثة من تقنيات الحماية الالكترونية و تهدف إلى كشف الحوادث المحتملة و المتعلقة باختراق إعدادات الوصول إلى البيانات في الوقت المناسب، و من ثم تطبيق سياسات الحماية بصورة تلقائية بغرض الحؤول دون اكتمال عملية التسريب. منع تسرب البيانات تعتبر من الاشياء المهمة في بيئات العمل و ذلك من أجل الحد من التأثير السلبي على الأعمال التجارية.

تستطيع أنظمة (منع تسرب البيانات) مراقبة البيانات في مختلف مراحل دورة حياة المعلومات وتطبيق سياسات للحد من المخاطر المرتبطة بها.

ويمكن استخدام تقنيات (منع تسرب البيانات) في المنظمات التي يحتاج فيها عدد كبير نسبيا من المستعملين النهائيين إلى الحصول على المعلومات ذات الطبيعة الحساسة. ويمكن أن تلعب دورا هاما في منع الكشف غير المصرح به عن هذه المعلومات وتخفيف التهديد المحتملة من داخل المنظمة. تقوم تقنيات (منع تسرب البيانات) برصد عمليات تبادل المعلومات وتطبيق سياسات محددة ضد الحوادث المشبوهة.

تكنولوجيا منع تسرب البيانات تعاني من بعض المعيقات التقنية و التي يمكن أن تقلل من الدور المطلوب منها في حماية البياناتز هذه الصعوبات يمكن أن تلخص في التالي: (1) صعوبات في وصف المعلومات بدقة ليتم رصدها، (2) المدة الزمنية الطويلة لتطبيق النظام، و (3) القيود في التعامل مع المعلومات المشفرة و الصور.

ويمكن تحسين تكنولوجيا منع تسرب البيانات عن طريق تعزيز أنشطة الرصد وسياساته بحيث يتم تحقيق إيجابيات أقل كاذبة والاستجابة في الوقت المناسب. تحسين تقنيات (منع تسرب البيانات) هو ضرورة في البيئات التي يحتاج امستخدمين فيها للتعامل مع كمية كبيرة من المعلومات الحساسة بانتظام. لان الوضع الراهن سسينتج عنه الكثير من (التنبيهات) الخاطئة و بالتالي سيتم تجاهل التنبيهات التي أثارها (منع تسرب البيانات) مما سيقود إلى شعور زائف بالحماية.

في هذا البحث، نقترح طرقا لتحسين تقنيات (منع تسرب البيانات) في شركات النفط والغاز من خلال تعزيز تحديد المحتوى الحساس في الرصد بحيث ينتج عن ذلك أقل عدد من التنبيهات الكاذبة. سيساعد تحقيق ذلك جميع المنظمات في هذا المجال على تحقيق التوازن بين أهداف حماية المعلومات والحفاظ على الوفاء بالحاجة المستمرة إلى تبادل المعلومات

CHAPTER 1

INTRODUCTION

In today's world, data is the lifeblood of any organization. Typically, data is acquired, processed, stored and interchanged on daily basis at records rates [1]. Maintaining data confidentiality, integrity and availability is a fundamental business need and must be fulfilled to maintain business continuity and to avoid devastating damage. Modern information security measures [10] are implemented in multi-layered models. Every layer contributes to the overall objective of safeguarding sensitive information within the organization. Figure [1] shows a typical implementation structure of information security measures within a given business domain:

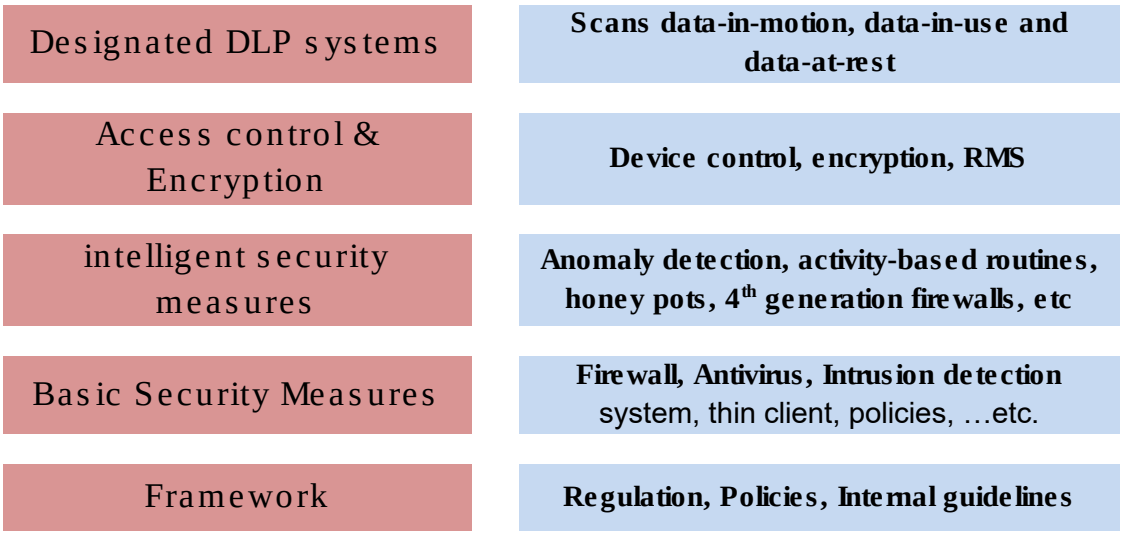

Figure 1: Structure of Information Security Layers.

Data Leakage Prevention is a relatively new technology. It is very useful in situations where access to sensitive data is always needed by end-users. Typically, Oil & Gas industry has this nature of data access. With the level of false-positives, complex configuration, and resources allocation of current Data Leakage Prevention (DLP) systems, organizations within Oil & Gas industry have one of two options: (1) Use technologies that are currently available and live with aforementioned high false-positive ratio and technical issues, and (2) Overlook the technology and give away a great chance to protect against unauthorized distribution of the organization's sensitive data.

In Oil & Gas industries, information confidentiality is a top business priority [9]. The vast majority of actions related to exploration, drilling, reservoir management, and production of Oil & Gas resources are nothing more than producing intellectual properties. Leaking such values comes with huge reputation and financial burden on the business. Therefore, many solutions such as Data Leakage Prevention (DLP) are implemented on top of security models to mitigate this risk.

DLP solutions can be classified into two main categories according to the way they detect sensitive content: rule-based or analysis-based. Rule-based DLPs [12] are discovering sensitive content using configured rules that define keywords, data-flow channels and other characteristics. Accordingly, the solution will assess the legitimacy of the data flow and trigger related action(s). Analysis-based DLPs use artificial intelligence or other analysis techniques to identify sensitive content and

create rules for near-matches. Rule-based DLPs suffer from high false-positives [14, 18]. Therefore, it will not be a good solution for Oil & Gas industry since the nature of the business in this domain relies heavily on timely and complete interchange of sensitive data. On the other hand, analysis-based DLPs, like text classifiers will provide an advantage over rule-based DLPs since majority of data within this domain is sensitive. As stated by [8], analysis techniques are also likely to fail in detecting documents, where most of the document is not confidential. This scenario is not applicable to Oil and Gas industries where the vast majority of interchanged data is technical report that are confidential by nature.

Once an organization commits to deploy a DLP solution, data needs to be classified. DLP solutions provided by lead IT vendors rely on manual data classification while Oil and Gas business requires that data classification must be done automatically so that human errors and malicious intent can be eliminated [13].

In this research, we argue that rule-based DLP solution can be improved drastically by devising a model that employs a comprehensive and structured analysis stage for the content before handling it by the used DLP solution.

Our proposed model provides the ability to automatically identify and classify sensitive content without any manual tagging or preprocessing. The proposed DLP solution is suitable for Oil & Gas and uses (a) Domain-specific taxonomy to support identification of sensitive content, (b) Text classifier to classify documents based on

the taxonomy attributes, and (c) File labeler to update the file header with taxonomy values.

The proposed model will improve DLP in the domain by reducing the number of false-positives. It will also improve performance by distributing the processing overhead of subject file over multiple stages. Our proposal includes a domain-specific taxonomy that will provide a standard categorization scheme. The taxonomy will be used by classifier to generate meta-data for each file; later on the file labeler will use this metadata to tag files. The applied DLP solution will use updated file tags and file timestamp to decide on appropriate action.

1.1 Thesis Objectives

This research will focus on using text classifier and domain specific taxonomy to identify sensitive content in a standardized and consistent way. DLP system used in the organization can be easily configured and operated with minimal resources allocation to identify sensitive content and respond accordingly.

The research aims to successfully reduce number of false-positives in files used in Oil & Gas domain and to reduce the time and resources needed to use DLP systems.

The research will also propose domain-specific information taxonomy to replace human input as the reference in classifying sensitive content. Reducing human factor in the classification process should result in higher consistency and standardization in the discovery of sensitive content.

The research will use domain corpora to configure the used classifier, conduct the experiment and evaluate results.

1.2 Problem Statement

The number of false-positives in DLP technology implementation within Oil & Gas industry domain is extremely high due inaccuracy in identifying sensitive content as well as the high demand on sensitive data by end users.

This can be improved by devising a model that incorporates domain-specific taxonomy and machine learning algorithms to learn what is sensitive and classify files in a complete automatic manner.

This model is expected to reduce the implementation time of DLP solution drastically and help organizations to improve the protection of sensitive information in the work domain. It will also avoid unnecessary delays and avoid hindering operations by monitoring data throughout the organization network. Accordingly, DLP solution can properly rank incident severity so that adequate mitigation can be applied to protect against information leakage.

1.3 Thesis Contributions

The proposed solution is expected to:

- Provide improved accuracy in DLP systems
- Reduce the time required to configure and maintain DLP technologies

- Introduce a new taxonomy for information within Oil & Gas industry

We will examine the ratio of false-positive upon using the enhanced model for DLP and compare it to the ratio of false-positive of a standard DLP system. We assume that the ratio should be much less after using the enhanced model.

In addition to that, we will discuss the requirement to implement the proposed model in comparison to the implementation of standard DLP system. We assume that implementation of the proposed model should be much less time and effort than standard DLP system. If both assumptions are found to be true, then we can conclude that devising DLP systems with automatic text classification and domain-specific taxonomy for Oil & Gas industry will improve performance and reduce overhead.

1.4 Thesis Outline

The thesis is organized into six Chapters. The first Chapter will provide an introduction about data protection challenges that different organizations face. In particular, the Chapter highlights this issue in Oil & Gas industry and relates the nature of business to the required contribution from DLP solutions. The Chapter also identifies major difficulties that might prevent organizations from reaping the fruit of DLP like the extended difficulties in accurately identifying information to be monitored and the lack of standardized way to configure and deploy standards DLP solutions. In the same Chapter, we state the research question and the method

to be used in order to resolve it using a combination of text-classification and domain-specific taxonomy. The Chapter ends with a brief of the expected results and contribution of the research and the method used to validate results.

The second Chapter shows a literature review of Data Leakage Prevention technologies at three different dimensions. The first one we reviewed the scale of the problem and expected role from DLP system in the business. We also reviewed the different suggestions by other researchers to improve DLP systems. The Chapter also includes quick description of data protection solutions other than DLP systems.

In the third Chapter we show in-depth discussion of DLP components, models, and architecture. In addition to that, it shows in details the different techniques used for content analysis in DLP solutions. We also used this Chapter to present information about the use of information in Oil & Gas industry. This information will provide necessary background to develop the proposed domain-specific taxonomy and its structure.

In the fourth Chapter, we introduce our proposed model, its structure, and its functions. We provide detailed information about the different components in the model and their assigned tasks. We also provide detailed information about the proposed information taxonomy and its uses in the proposed model.

In the fifth Chapter we provide information about the implemented prototype including the text classification model and the information taxonomy model. The

Chapter provides details about the used datasets and conducted experiment. Results of the experiment will also be provided in this Chapter to decide on the performance of the prototype.

Chapter six concludes this thesis with a summary of the objectives, analysis, proposed model, performance evaluation and experiment results. We also envision areas open for improvement and envision future extensions.

CHAPTER 2

LITERATURE REVIEW

DLP is not researched highly in the academic community. The technology is still building up and there is a need for more researches and development. Recently, more researchers are looking for improvement in DLP technologies. Researches are expected to reflect positively on the role and the contribution of DLP in the protection against unauthorized distribution of sensitive data.

In this Chapter, we will provide information about research work and technologies in three main areas: Scales of the problem, DLP researches, and non-DLP Data Protection solutions.

2.1 Data Leakage Problem: Scale and Occurrence

Unauthorized distribution of sensitive information within the enterprise network is emerging as an ongoing challenge to all organizations. Such distribution will result in an adverse impact on the organization business and reputation. It is the sole responsibility of the organization to properly handle confidential data under its custody. A shortcoming in this domain is a violation to the government regulations and might lead to business loss, government fines, and law suits.

Protecting against cyber-security threats and other data loss risk is a direct responsibility of the organization's CEO and its board. Introducing proper internal controls is a compliance requirement. Criminal penalties for certain misconduct in this domain are enforced by the name of law.

The financial impact of data loss has been studied by many researchers. According to Verizon [41] data breaches is steadily increasing on the last five years. One has suggested that addressing this issue the decade challenge for computer science.

According to the Japan Network Security Association, sensitive data loss incidents in 2015 were 864 affecting more that 30 million people. The total damage is estimated to be more than 21 million US dollars [34]. These incidents are taking place in a regular basis as a result on intentional or unintentional actions. It has been reported that 66.2% of business users have sent emails erroneously [35]. Datalossdb, a nonprofit organization that documents known and reported data loss incidents world-wide, reported a constantly increasing number of data breaches as show in Figure 2.

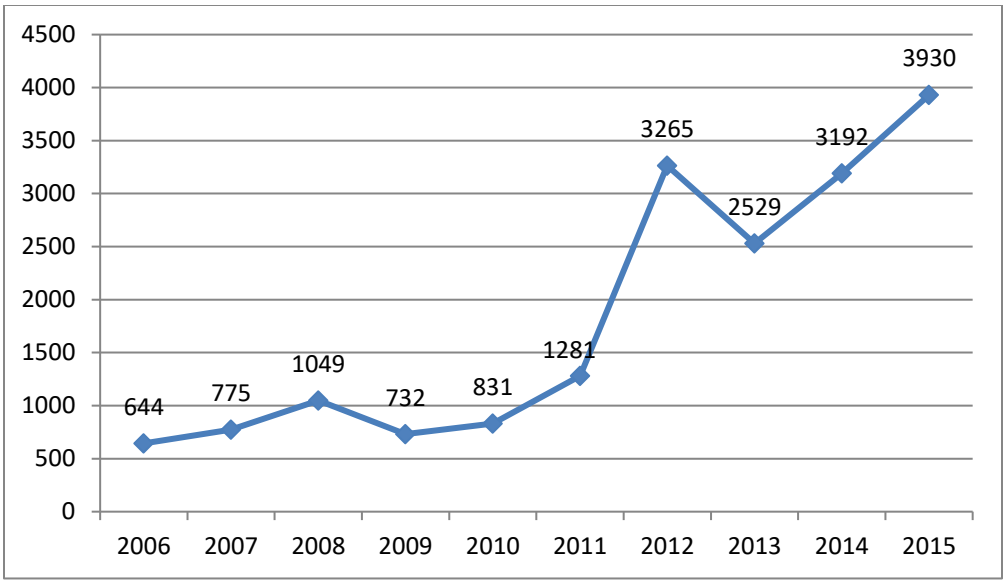

Figure 2: Number of data breach Incidents by datalossdb organization.

Clearinghouse, another nonprofit consumer organization, reported that 227,052,199 sensitive records that reveal personal information were leaked during the period from 2005 to 2014.

The very obvious example of the financial and reputational damage that happen to organization due to data leakage is the consequences of content published in WikiLeaks. Almost all information published there are classified and shared by insiders who made it through traditional protection measures.

The global Information Systems Security report for year 2016-2017 showed that Oil & Gas industry is improving their understanding and preparation to protect against data leakage as part of cyber security. However the report stated that there is a need among the industry for improved protection to potential leakage of sensitive data.

Although [8] has argued that company reputation cost is overstated; [7] predicts that loss of reputation is one consequence of a data loss. Robert et al. had studied a limited sample that doesn't represent the actual interest for people like social media network. His research didn't cover critical sectors like banking, stocks, monetary firms or major manufacturer. His sample was limited to only two case studies in which one of them is a social network company.

Different methods were suggested to estimate the cost of data leakage. Bunker suggested that loss can be divided into two categories: Direct loss: tangible damage can be measured like fines due to regulatory breach, clients compensations, law suits, forensics cost, loss of man-hours, restoration cost. Indirect cost: stock share, reputation, loss of customers, loss of investment, loss of intellectual property and trade secrets.

Unauthorized distribution of trade secrets and personal information within organization content can results in a serious damage to the organization. DLP can be used to protect against such consequences and maintain the organizations in control over their sensitive information. DLP protection covers: (a) Data in motion: identify and block unauthorized information transfer, (b) Data at rest: identify and proactively protect data in different repositories, and (c) Data in use: apply protection measure to protect against authorization abuse.

Many DLP solutions are available in the market from major vendors in information security sector like Symantec, McAfee, RSA, and others. However, the expected

role of DLP systems is not yet properly addressed by current technology [30]. The main reason for that is the difficulty in identifying sensitive content. Comparative study to assess the relevancy of data leakage prevention mechanisms [30] revealed that DLP technologies need to improve mainly on the way sensitive data is detected and protected from unauthorized distribution. The researcher highlighted that DLP solutions need to improve on the semantic rather than syntax since structure of information might change from one document to another while same sensitivity is carried by two representations. There should be ways where a deployed DLP solution can identify “Information” instead of identifying the “Template”.

2.2 Improving DLP Systems

Data Leakage prevention is considered to be a new dimension in information protection and expected to protect against insider threats. DLP emerged as a result of repeated data leakage incidents and the regulatory compliance requirements. Researchers have looked at different areas of interest to improve DLP systems. In this section, we will provide a survey of researchers work.

Many researchers have argued that DLP systems can be improved by focusing on the semantics and context rather than syntax and structure of monitored content. They also argued that results from DLP system are extremely impacted by data format and repositories. Preeti Raman et al. [30] for example concluded that suitable prevention techniques are defined by the repository in use. He also argued that DLP

systems need to consider the context in order to infer the semantics of communication. Other researchers have also emphasized on the importance of the semantics to improve DLP system when handling a given data leakage incident. Sandip et al. [39] emphasized on the semantics of the event in addition to data. In other words, he was highlighting the context in which data interchange event is taking place. He argued that lacking semantics is a common mistake in current DLP solutions. He suggested that future researches should addresses this area.

Other researchers suggested different techniques to improve DLP systems. For example, Vachharajani et al. [31] suggested a labeling mechanism in which data will be assigned pre-defined tags to ensure that monitoring components will be able to capture suspicious incidents and triggers mapped policies. He did propose a method to perform this tagging mechanism. Some implementation of DLP systems requires that labeling will be performed by end-users which will open the door for high inaccuracy and malicious intent. Radwan et al. [40] suggested working on developing text classifiers and integrating encryption systems as a future work on Data Leakage Prevention systems. He argued that such work will solve the problems with current systems.

Some researchers argued that DLP performance can be improved by matching the technology used in discovering sensitive content with the implementation environment. They classified DLP discovery technologies into rule-based and analysis-based [14, 18]. They argued that rule-based DLP systems suffer from high

false-positives; hence, analysis-based DLPs will provide better results. On the same direction, [8] stated that analysis-based DLPs are also likely to fail in detecting documents, where most of the document is not confidential. We could not find a research in the literature that suggests a combination of both techniques in order to complement each other and provide higher performance results.

Other researches discussed that DLP systems can be improved by enhancing the processing overhead. Given that DLP systems will handle huge amount of data, it becomes very necessary to balance processing needs in order to avoid latency/disruption to day to day activities [50]. Separating DLP activities between end-point and DLP gateway will help reduce performance overhead.

Some work in the literature discussed methods that help investigation of leakage incidents. Several ways have been suggested to support effort that takes place after leakage is encountered.

Some researchers suggested a method to gather Operating System files that help in investigating data leakage incidents. Lee et al. [32] provided guidance in files to be collected for windows based platforms.

On the same direction, White et al. [39] proposed another method that helps identifying the leakage source by generating uniquely identifiable records that looks real so that forensics can identify the source of the leakage.

One cannot decide on the leakage incident without a comprehensive understanding of the domain. Author in [30] concluded that full knowledge of the context and the patterns in which data is used are necessary elements to imply if data leakage incident is taking place or not.

Rohit Pol et al. [38] discussed the likelihood that an agent is responsible for a leak is assessed based on the overlap of his data with the leaked data. They presented an algorithm to implement a variety of data distribution strategies.

Some researchers argued that DLP solution by itself cannot protect against data leakage. Tomoyoshi Takaebayashi et al. concluded that DLP should include three different yet integrated solutions like USB protection, email filtering, and secure document management solutions in order to properly address risk of data leakage at different data stages.

Other researchers also emphasized on the subject of integrating DLP solutions with other information security measures. Ernst & Young's Advisory Services [48] stated that effective protection of business sensitive data must be handled as a program rather than a solution to be deployed. The report identified data classification, user provisioning, regulatory framework, risk-assessment methodology, incident response measure as integral components to such a program.

2.3 Non DLP Data Protection Methods

Researchers have identified different ways and methods for data protection other than DLP solutions. Most of these ways are focusing on protection against threat from outside the network like hackers or intruders.

2.3.1 IDS / IPS

The main objective of Intrusion Detection Systems (IDS) is to monitor the traffic in the network and the activities in attached systems in order to identify cases where malicious activities are suspected. IDS come as hardware and in other implementation are built as a software component within other network devices.

Intrusion Prevention Systems (IPS) performs almost same functionality as IPS. It monitors the traffic in the network and configured systems in order to identify malicious activities. In addition to that, IPS is designed to defuse suspicious activities and keeps information for future use. In that sense, we look at IPS as an extension to IDS.

IDS/IPS typically includes hardware and software components that work together to complete the whole task. Usually following components are found in such systems: (a) Detectors: Capture data in the network and send it to the central system. (b) Central system: Receives captured data and apply necessary analysis. (c) Business intelligence and reporting: Generate reports based on applied analysis. (d) Database: Stores information needed for analysis and protection like source of

previous attacks, trend analysis, etc. (e) Response container: Collects and store information and form appropriate responses.

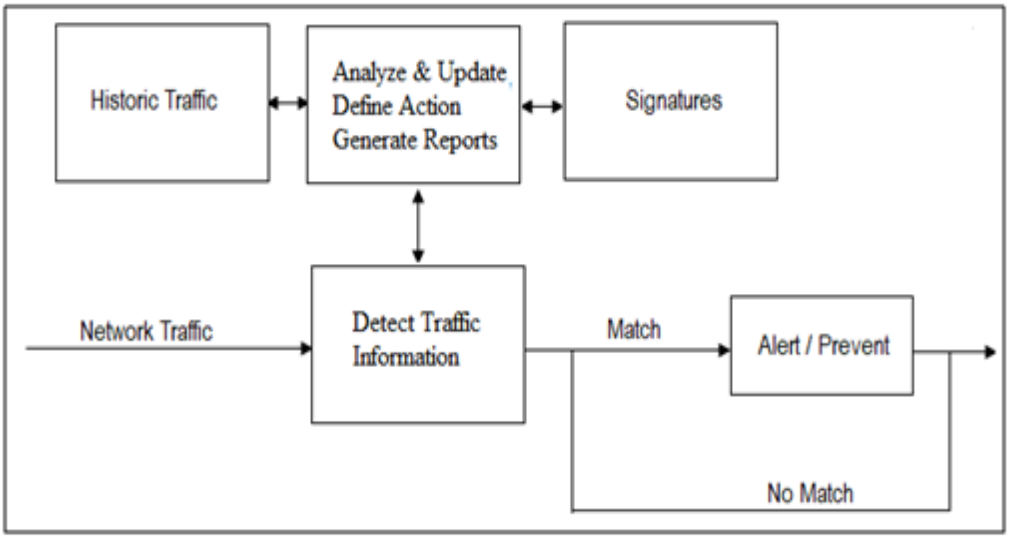

Figure 3: IDS/IPS system based on Anomaly Detection.

IDS / IPS follow two main techniques in monitoring traffic: (a) signature-based or matching (b) Heuristic or rules. Signature-based approach uses either pattern matching or packet classification to identify an occurrence of intrusion trial based on recorded attacks. A database is used to store information about viruses, worms, Trojans, and other malware to compare it with suspected intrusion trials. The next diagram shows typical components of Signature-Based systems [30]. Heuristic or rules based systems monitor the activities in the enterprise network in order to identify up-normal situations.

These systems are very efficient in identifying anomalies and known patterns; however, they are not as efficient in identifying isolated incidents of leaking sensitive data outside the network.

2.3.2 Anti-Malware

Malware is a malicious code/software that aims to cause damage in computer activities, steal sensitive information, modify or delete data values, etc. There are many types of malwares like viruses, worms, Trojan horses, spywares, embedded malicious code, ransomware, rootkits, etc.

Anti-malware is any software that recognizes and removes malware from computer devices. It also blocks installation of infected components (i.e. files, Trojans, etc.) in the computer. Anti-malware is an efficient protection tool and stand as a necessity to protect against threats from outside the network. It validates the inbound traffic using databases of known malwares to ensure that malicious content will not be allowed in. Anti-malware solution can protect two types of threats: Insider threat, zero-days attacks. They should be bundled with other protection technologies to improve the readiness of the network and narrow chances of successful attack. However, antimalware cannot protect against abuse of authorization. Data leaked through authorized accounts can go unnoticed by these protection solutions.

2.3.3 Firewalls

These are the basic defense mechanism to control the flow of network traffic to and outside the network. Firewalls are basically used to configure the set of accepted

access rules and to identify unauthorized access. Firewalls technology has evolved into Next-Generation Firewall (NGFW) that combines other functionality with access control. Currently, NGFW comes bundled all together with Intrusion Prevention, Deep Packet Inspection, and Anti-malware capabilities [27].

Firewalls can protect against viruses, worms, malware, and Trojan horses. Without a firewall, the network becomes open for attacks from outside which might result in loosing sensitive data in different repositories within the network. However, data leakage that is originated from within the network can not benefit a lot from these capabilities. It is necessary to integrate firewalls with other information security to protect against such risk.

2.3.4 Others

There are many other solutions to improve protection of sensitive data. The number and type of solution has exponentially increased as a result of the expansion in Information and the need for protection solution. Multiple layers in the information security framework as shown in Figure 4 can be used to enhance the protection of sensitive data in a given network.

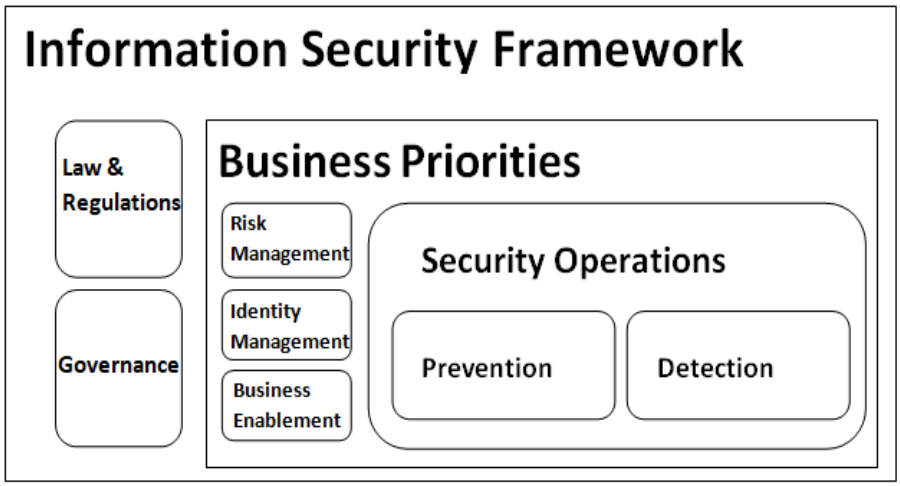

Figure 4: Information Security sub-domains.

Data protection can be achieved by other tools that address one or more protection dimension. An example is the use of Google Alert as a replacement for the Discovery function of DLP systems. Google Alert can monitor web servers and automatically notify concerned organization about availability of sensitive data in a certain domain so that protection measures can be applied.

Some Web Security and Email protection solutions includes data leakage as part of their functionality. Webense WebGate Security has developed PortAuthority solution to include some leakage prevention capabilities.

Other web blocking solutions like HTTP/HTTPS blockers, SMTP blocking - FTP prevention server can be used to protect data by controlling the ability to transfer content and/or files to servers outside the network.

Deployed solutions in each sub-domain including DLP systems can be configured in Security Information & Event Management (SIEM) system in order to provide

data and analyze it. The goal of this integration is to highlight vulnerabilities and help getting a centric visibility to the possible threats on the network [33].

CHAPTER 3

DLP Systems and Research Domain

Data Leakage Prevention systems includes several components that work together to complete the intended actions. It is important to have solid background about all of these components, their functions, and their architecture in order to propose new model. It is also important to understand the techniques used in identifying sensitive content in today's Data Leakage Prevention systems so that we can compare it and maintain the capabilities in the proposed mode.

It is also in the same importance to have a solid background about the use of information within Oil & Gas industry. As a targeted domain by our research, we should have clear idea about data capturing, data flows, and data storage activities. We also need to have good idea about business functions and the expected data consumption at different stages so that we can estimate the importance and the criticality of devising a reliable solution in such a domain.

In the following sections, we will discuss the components and the working model of Data Leakage Prevention systems. We will also discuss in details the content analysis methods used to identify sensitive content. Our discussion will provide the required background about strengths and weaknesses of each method so that we can proceed to propose the suggested model. We will also discuss the use of information

in Oil & Gas domain showing their nature and the common activities related to this issue.

3.1 Analysis of DLP Technology

DLP system has different components that work together and integrate each other. They are built and implemented on the enterprise network in a way that facilitates the required function. DLP solutions use different techniques to complete content analysis as part of the overall function. In the following section we will discuss the structure, function, and techniques used in DLP systems.

3.1.1 DLP Components

DLP has two main broad components. Together, they enable monitoring and preventing of unauthorized distribution of sensitive information. These components are: (a) Data Monitoring Components and (b) DLP policies components.

Data Monitoring Components handle the capture, rebuild, and analysis of transferred files. The techniques to complete this task are dependent on the file type, file format and file location. Together, they got the file collected, rebuilt, and passed to the “Content Analysis” part in order to decide whether an existing policy can be associated with it or not. Content analysis is the main component in DLP systems. It evaluates the content and specifies the correct policy to trigger. We will discuss content analysis in more details in the following pages.

DLP Policies components contain information about: (a) Repositories: define the location where data to be protected is stored. (b) Rules: defines when and how to protect content. Two things are taken into account while working with repositories and rules: 1) the conditions to be matched. For example, look for content with Oil well number and production volume, and 2) the action to be applies: Automatic response when the condition is matched, for example, takes a log and report the event to some recipients

3.1.2 DLP systems Architecture

DLP is a multi-layered solution. Once the first layer of protection is implemented, the next layer should/could be addressed. There are many different forms of DLP applications architecture depending on the location of the sensitive data and the transfer method. A typical DLP system architecture is shown in Figure (5)

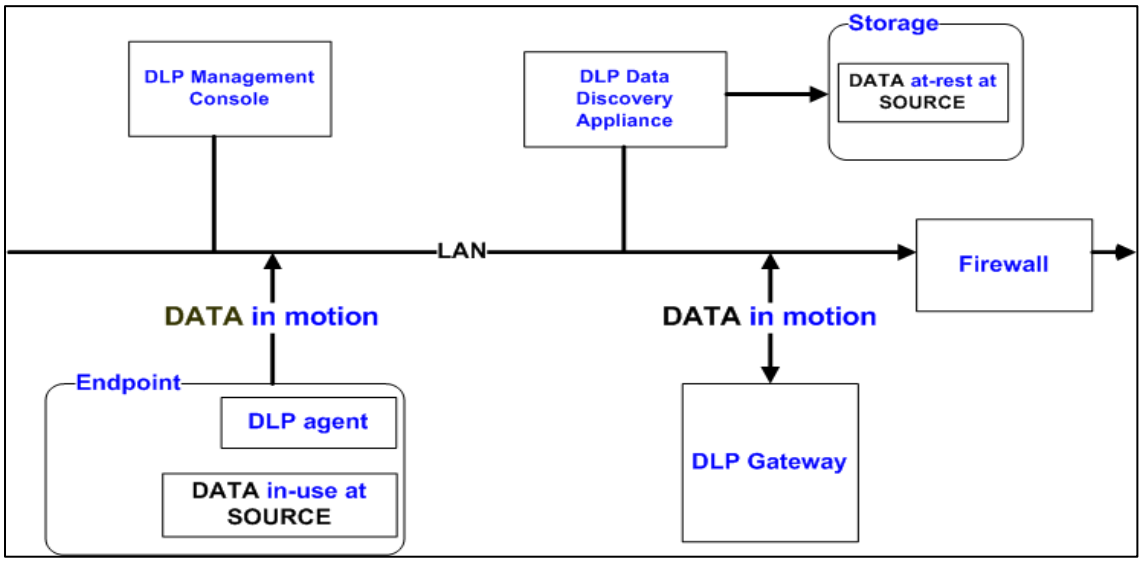

Figure 5: DLP System Overall Architecture.

Data Monitoring and Analysis tools usually architecture in a way that facilitates capturing of data. Typically, they either exist at data repository site (i.e. end-points) or as a separate appliance to intercept data transferred over the network. DLP Policies components usually exist in the central management site. They facilitate centralized responses/actions to different sensitive data transfer scenarios.

Data at source and data at rest are two different statuses in which data resides in end-use computer or other storage station (i.e. file server, NAS, etc.). DLP systems can handle data in this status and trigger actions like blocking movement of data or generate an alert. Data in motion is the third status in which DLP can deal with data, typically, DLP systems can block or allow the transfer.

Data that is stored in different repositories will be accessed and scanned by DLP system in order to identify if a sensitive content is available or not. Accessing these files require network administrator to define two things: (a) Administrative connection for DLP agents in order to traverse through the network and reach for files. (b) Central DLP policy/ scanning server.

This architecture will carry high-traffic in the network thus; enough resources are needed to perform real-time analysis and scanning. The other option is to complete the scan after peak-hours which open the window for leaking data during working hours (the actual time where leakage might happen). Another way to work with files that resides in different repositories is deploying a server agent module in the repository. This way is used in situations where huge amount of data is stored in

one repository such as database servers and file servers. The agent will perform required scanning and send results and findings to the central DLP server. Network and database administrators will be required to monitor the impact of these agents on these servers. Another architecture of DLP systems is to integrate with other solutions that manage documents such *Documentum* (*enterprise content management platform*), *FileNet* (software to help enterprises manage their content and business processes). Integration is needed in these situations in order to perform all discovery and analysis tasks.

An agent will be installed in the server side and configured in a way that it can reference storage attributes, and documents attributes in a step to perform the scanning and analysis tasks. Configuration of such agents requires multiple tests and usually requires second line support in order to match changes on the targeted solutions.

3.1.3 Content Analysis in DLP Systems

Content Analysis is actually a collection of techniques used to understand the monitored information. It can be divided into two broad areas: (a) Described data: Specifying how to identify a piece of information like “Well Information”, “Reservoir Information”, “Production Information”, etc. and how it looks like (b) Registered data: Specific document or specific database or sub division of either one

Content analysis in DLP systems sometimes is referred to as Deep Content Inspection (DCI). It is considered as the evolution of Deep Packet Inspection and adds to the picture the ability to examine the whole file and its content rather than individual packets. DCI allows for a content-aware inspection. It rebuilds the file and look for matches as defined in DLP policies. Figure 6 shows the architecture of Deep Content Inspection (DCI) [35].

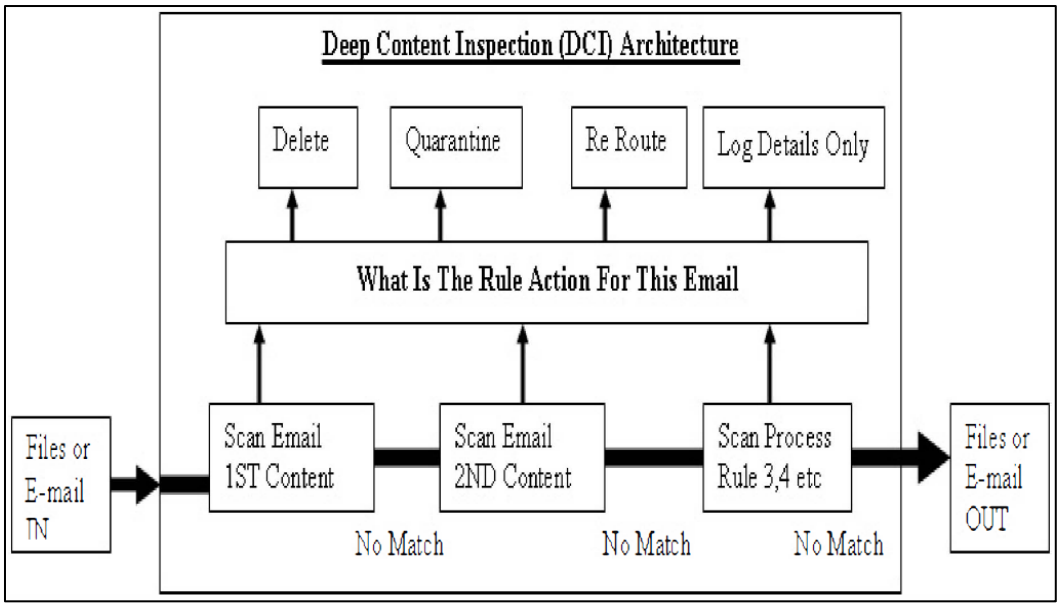

Figure 6: Architecture of Deep Content Inspection.

DCI is built using two main components: (1) Content Capturing & Collector (2) Content Analysis Engine.

Content Capturing & Collector parses textual data of a source file to pass into the content analysis engine. The analysis engines need to work with text, and many of the file and data formats such as Office documents or PDF files are binary data. The "Content Capture" takes a file, determines the format, then uses a parser to extract all the text.

Some tools can support hundreds of file types, including complex situations like documents embedded in other documents. It's the job of the collector to assemble the file and pass it for analysis

Once the file is opened up, the content analysis engine evaluates the text and looks for policy matches. On occasion, the tools will look for a binary, as opposed to a textual match, for data like audio and video files, but the textual analysis is where the real challenge is. Several techniques are used for this purpose like rules/regular expressions that uses textual analysis to find matching patterns, such as the structure of a credit card or Social Security number. Some of these rules and regular expressions can be quite complex to minimize false-positives. Another technique is database fingerprinting (exact data matching) pulls data from a database and looks only for matches of specified data. Thus, it is loaded with hash values. Partial document matching is another techniques that takes a source file, parses out the text and then looks for subsets of that text. It usually creates a series of overlapping hashes that allow capabilities like identify a single paragraph cut out of a protected document and pasted into a Web mail session. The latest used technique is the statistical analysis. It uses machine learning or other techniques to analyze a set of known "protected" data and known "clean" data to create rules for near-matches.

3.2 Information in Oil & Gas Industry

Oil and Gas play a very important role in the world's economy. The industrial nature of today's life places an increasing demand on hydrocarbon resources. In the early twentieth century, hydrocarbon resources became the most valuable commodity traded in the world markets [11]. The industry has three main sub-domains: (a) Upstream, (b) Midstream, and (c) Downstream. This research focuses on the first domain, the Upstream as it represents the most critical part in the industry and involves the most sensitive information.

3.2.1 Main Upstream Functions

Upstream domain includes three distinctive yet interdependent phases namely: Exploration, Development, and Production.

In Exploration phase, geologists and geophysicists work together in a quest to search for potential underground hydrocarbons. They collect information through seismic test, core samples and Drill Stem Test to identify rock properties (i.e. porosity, permeability), formation structure, flow rate, pressure, temperature and other data to characterize and model the reservoir [2].

In Development phase, petroleum engineers will work to manage the identified reservoir in a way that achieves the highest economic recovery. They accurately assess the reserves, recommend cost effective way to develop the field (i.e. well drilling and workover), production, and reservoir depletion. They use data from

geological models (Reservoir Description & simulation), production, pressure, temperature surveys to complete the field optimization task [2].

In Production phase, production engineers will monitor developed fields and perform surveys used in reservoir analysis [2]. In addition to that, engineers contribute to other functions that fall under different phases. For example, production engineers will participate in data acquisition data acquisition for the use of functions under Exploration and/or Development phases.

3.2.2 Data in Upstream Functions

There is a huge increase in data types and volumes within Oil & gas industry. All companies are facing difficulties in managing this ever-increasing data. Main challenges laying on identifying and implementing best techniques in order to store, quality check, retrieve, and protecting this data. Such difficulties have resulted in unnecessary delays in completing related business tasks. Specialists will spend longer times to get required information, and prepare it for their use. On the other hand, Management are frustrated due to extended cycle times. On the other hand, concerns related to uncontrolled access to data and data leakage is constantly increasing. The need to maintain the organization's knowledge and to introduce and maintain rigorous audit trail is also constantly increasing. In brief, Information management is becoming a problem for everyone. In many instances, these factors limit the organization capacity and result in increased expenditure. An effective way of managing information is having it categorized into functional and operational

categories. Knowing the nature of the information and its use in the business along with other identifying attributes will enable better management and better security.

For instance, if information is classified as confidential information that is applicable in certain business function, then there will be no worry about having this information interchanged over multiple steps or different processes within that function. Alerts should be raised when that information is handled by non-related process.

In Oil & Gas Industry, data is typically generated in one business function and handed to the next function to continue the value-chain. Identifying the business function that generated data and the destination or possible uses of this data is very important to deem the legitimacy of any data access activity. In addition to that, it is of the same significant to identify other factors related the data format/representation and the sensitivity attribute. A typical flow of data through different business functions within Oil & Gas domain is shown in Figure 7.

Sensitivity of information is directly related to the phase of generating and consuming data. For instance, data related to new explorations and field development is of a very high sensitivity. Information related to production volumes are of the same sensitivity.

Based on this notion, we suggest that information taxonomy for Oil and Gas is very important. It will be used to identify the sensitivity of a given content taking into account the domain-specific input (i.e. business function).

There are limited trials to build such taxonomy however, we are not aware of any completed effort that can be used for protecting sensitive data. Available taxonomies are considering either a high level classification drivers or absolute business functions. The useful taxonomy should provide all attribute needed for the goal of protecting the data. Hence, we need to develop a taxonomy that serves the research purpose and provide a step in standardizing information management activities including information security.

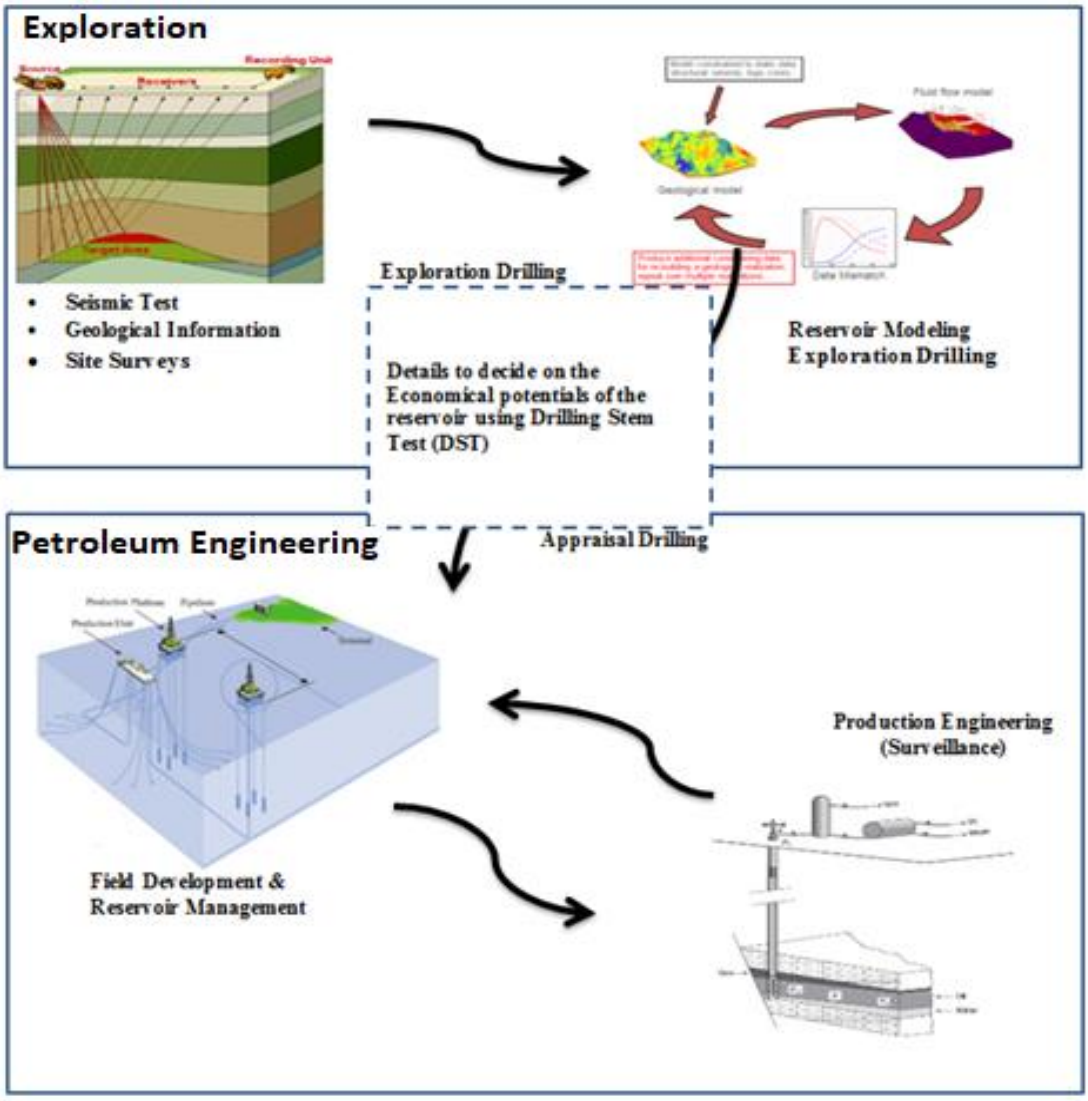

Figure 7: Information flow in Oil & Gas Processes [52].

The taxonomy will also be useful for other data management activities like quality assurance, knowledge management, and data analysis. It should provide the ability to store information related to these activities as well as information security activities. Studying different business functions and data generated / used in each function will help develop this taxonomy. Sample of different business function and data types (i.e. names) used in each function is shown in Table 1. The table shows data types into categories based on major business function in each phase.

Business function	Data Category	Data Types
Exploration	Seismic Data	Seismic surveys & Acquisition
		Seismic processing
		Seismic interpretation
		Seismic acquisition crew morning reporting
		Seismic borehole activities
	Core Data	Conventional core analysis
		Special Core Analysis
		Core storage
		Hydrocarbon shows
	Log Data	Open hole logs
		Cased hole logs
	Hydrology	Water well monitoring
	Geochemical Analysis	Geochemical Analysis
	Gravity and Magnetics	Gravity and magnetic surveys
	Stratigraphic Picks	Stratigraphic Picks
Production Management	Reservoir Basic Information	Fluid contact
		Formation test
		PVT data
		Formation testing and sampling (FTS)
		Reservoir pressure (SIBHP, PTA)
		Average reservoir pressure
	Production Management data	Annuli survey
		Avails
		Crude sample analysis
		ESP performance
		Flowmeter survey
		Gas injection
		Gas sample analysis
		Gas well monitoring
		Inhibitor jobs
		Oil production

Table 1: Sample of the data used in Oil & Gas industry.

The need to identify metadata necessary to enable better data management and improved security is mostly overlooked by generic Data Protection tools and has not been clearly stated in most of the previous implementations. In order to include this metadata in the proposed Data Leakage Protection model, it is needed to establish a standard way for classification. For this reason, we suggest that taxonomy of information in Oil & Gas industry need to be established.

CHAPTER 4

ENHANCED DLP PROPOSED MODEL

We suggest that a better solution can be achieved by reducing the ratio of false-positives and balancing the processing overhead. We introduced a model that works in the following logic:

1- A module will use Support Vector Machine (SVM) model to classify the files and update the file metadata based on the used information taxonomy. The updated metadata of files will be used later on for different purposes including security related goals.
2- The DLP solution will use the file metadata to identify sensitive content for data-in-motion and trigger the designated counter measure as defined in the solution policies.

The proposed model is structured in a way that enables DLP solution to perform required tasks with the minimum overhead while keeping other components perform in updated and synchronized manner. Files to be classified exist in end-users storage media, file servers, or other file repositories such as *Documentum*, *FileNet*, etc. each repository will subscribe to the service by registering identification information of the device and ensure that information taxonomy is up to date.

The proposed model scans files in the repository to identify obsolete timestamp. These files are classified using an SVM-based text classifier module and results will be provided to the “File Labeler” module. On the next step, the labeler updates the timestamp and the file label with identified values for each attribute in the used taxonomy based on results from the classifier.

Data-at-motion is examined by the deployed DLP solution (Rule-based) to identify files with specific labels. The DLP solution triggers adequate actions/policies accordingly.

The proposed model includes the following components:

- Information Taxonomy: Acts as a knowledgebase to provide details necessary to complete the classification of files within Oil & Gas domain.
- File Checker: Ensures that stored files are classified
- Text Classifier: Uses information stored in the taxonomy to classify and generate attributes values for stored files.
- File Labeler: Updates file with a label that reflects the identified attributes values.
- Corporate DLP: Rule-based DLP solution that is implemented on the domain network

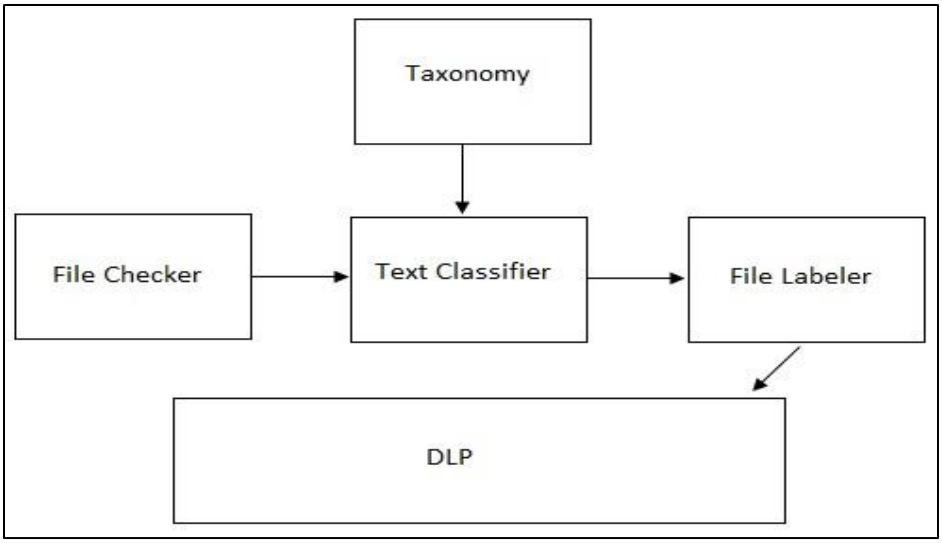

Figure 8: Block Diagram of the Proposed Model.

4.1 Functional Architecture

The different components will interact with each other to perform one task that is reducing the false-positives by having files classified and prepared in accurate, standardized and consistent manner. This will require files to do exist in physical repositories where other system components will access and classify them. Communication components will use the network to retrieve files and taxonomy information from their repositories. Retrieved information will be passed to other components in the model through queue systems in order to be classified and labeled.

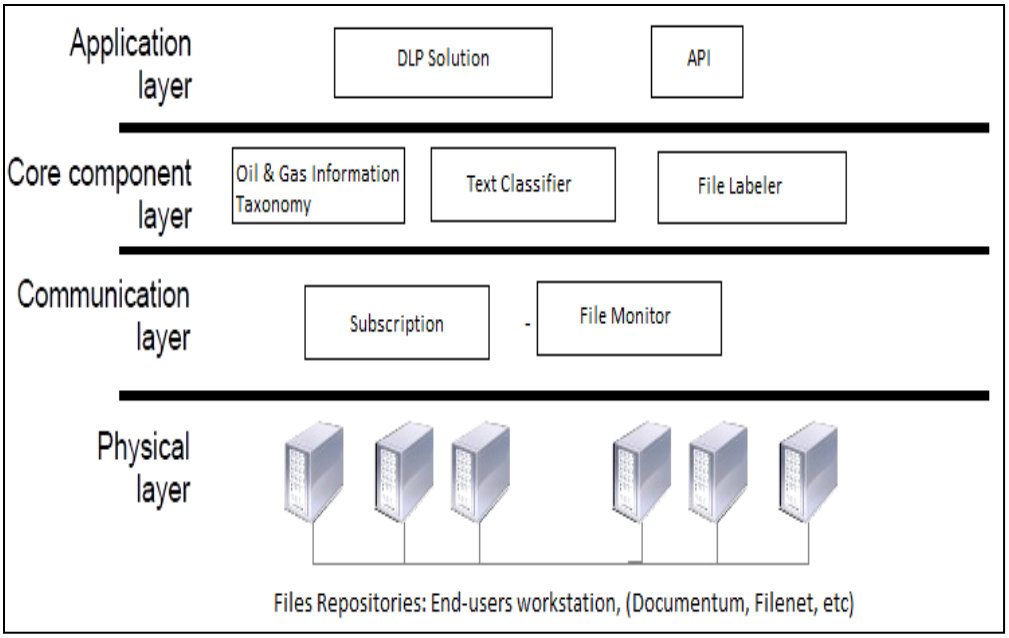

Figure 9: Functional Architecture of the proposed model.

The main components in the system are the Text Classifier and The File Labeler. The text classifier will get files in the queue and perform classification based on attributes defined in the local copy of the taxonomy. Identified attributes values will be used by the File Labeler to tag the file. The File Labeler will also update the timestamp to ensure that files have been classified after last change. These functions are working with each other as described in the Figure 9.

4.2 Operational Architecture

As stated by [14], an operational architecture consists of a set of software components, a set of flows between the components, and a set of constraints on the components. Our model operates in the way described in Figure 10

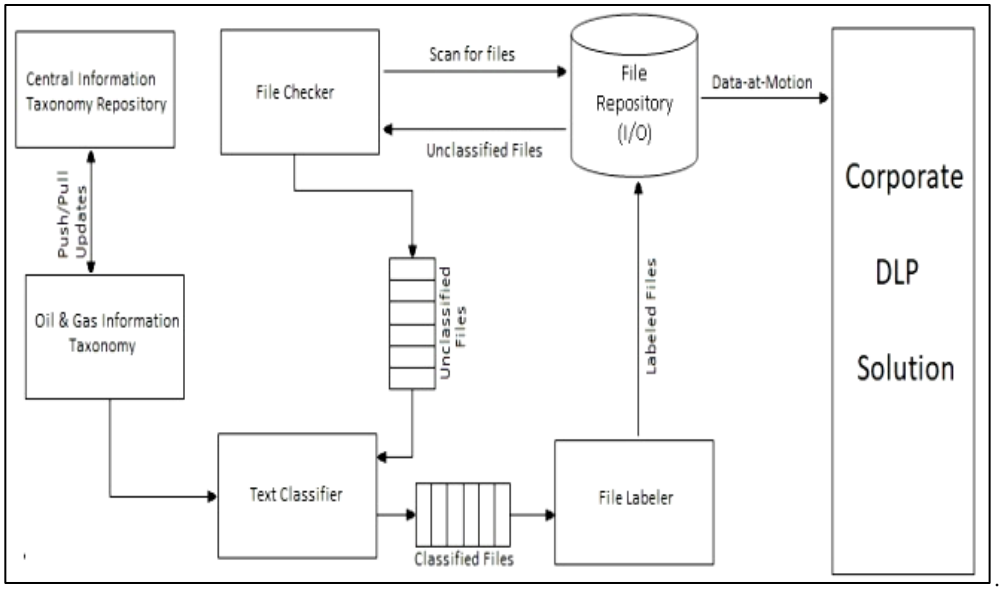

Figure 10: Operational Architecture of the proposed model.

The proposed model operates on files that exist in "File Repository". On cyclic checks, the "File Checker" will evaluate stored files to decide if classification is needed or not. File to be classified will be placed in a To-Do queue so that "Text Classifier" will handle them in ordered manner. Classification results will be sent to the "File Labeler" along with file information. The "File Labeler" will update metadata of each file with results received from the "Text Classifier" and write the file back to its original storage location.

When files are sent outside the network domain, they will be intercepted by the "Corporate DLP". Files will be received with clear, standardized, and consistent labels that enable for accurate (low false-positive ratio) actions to be executed.

4.3 Proposed model components

The proposed model has several components that interact with each other and with other systems to complete aimed enhancement. Following is a description of each component:

4.3.1 File Checker

The File Checker (FC) will scan files that exist on the designated repositories (R). It will look for the file classification-timestamp (f_{cts}) in each file and compare it with the file last change timestamp (f_{LTS}). Files with (f_{cts}) < (f_{LTS}) in (R) are the candidates for the Text Classifier. These files will be locked by the (FC) and placed in the Unclassified Files List (UFL).

File Identification for classification

(1) ; ; Scans files from designated repository (R)
(2) ; ; Add Unclassified files to UFL
(3) For each file $f \epsilon R$
(4) *Get* f_{cts}
(5) *Get* f_{LTS}
(6) *If* $(f_{cts} < f_{LTS})$ *then*
(7) Lock *(f)*
(8) UFL = $UFL \cup f$

Figure 11: File identification algorithm.

The File Checker performs the scan for text classification candidate files in pre-defined Check-Cycles (CC). The cycle interval can be configured in a way that doesn't overload the system or adversely impact the performance. The algorithm shown in Figure 11 describes the operation executed by the File Checker (FC) in order to identify candidate files for classification.

4.3.2 Oil & Gas Information Taxonomy

Oil & Gas companies are facing challenges in managing the ever-increasing data in the domain. Main challenges laying on identifying and implementing best techniques in order to store, quality check, retrieve, and protecting this data

We suggest that information taxonomy for Oil and Gas domain will improve the performance of data security. It will help identifying the sensitivity of different data types in a standardized and consistent way.

The proposed taxonomy includes several attributes that define the source, uses, responsibility, and sensitivity of information used in Oil & Gas industry. For every data set, the taxonomy value will be identified in a way that takes into account the value of each attribute of that data set. The identified value will be used to determine how to handle different aspects of data management activities including data security. The proposed taxonomy includes the following attributes:

(1) Geographical attributes:

This attributes define the exact geographical location to which data is related. Examples: Eastern Oil Fields, African Oil fields, Central Asian Oil Fields…etc. Geographical locations are usually organized in structural hierarchy that consists of one or multiple locations. For instance, African Oil Field may include subordinate of the following values: Nigerian fields, Angola Fields, etc. and the Arabia Gulf Fields may include the following subordinates: Saudi fields, Kuwait fields, Oman fields, etc.

The geographical attribute is designed in a way similar to the scheme used by Oil & Gas companies in segregating their operations. Usually, a level of classification is set for the main region (i.e. content) then the region in which hydrocarbon reservoirs might expand over multiple geopolitical locations (i.e. Arabian Gulf, Central Asia, etc.). The last level of classification is the Oil/Gas Field. The structure diagram for geographical attribute is shown in Figure 12. The figure also shows examples of the attribute values

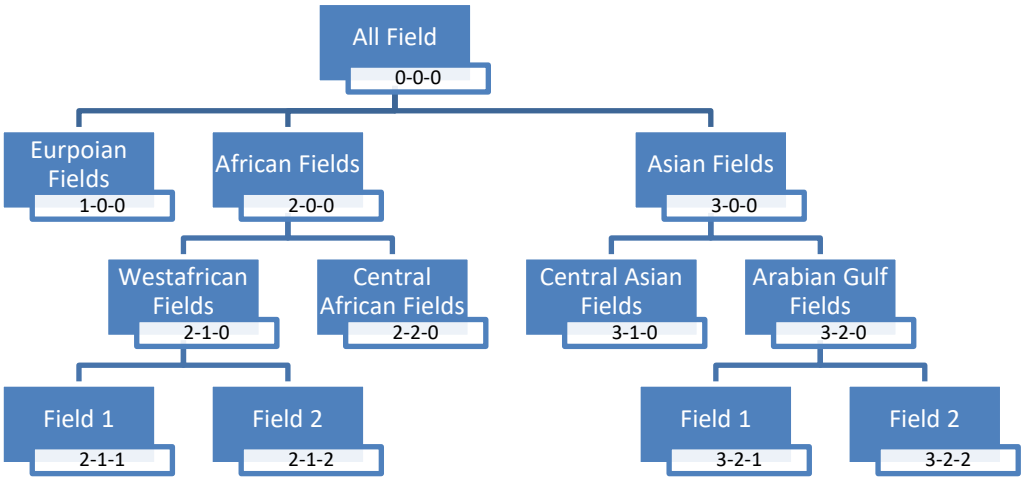

Figure 12: Taxonomy Geographical Attribute.

(2) Organizational attributes:

This attribute defines the organizational entity within the enterprise that generated this data. Usually, such organization is the main beneficial of this data and the most concerned with its quality and security. This attribute will be used to decide on the legitimacy of accessing and handling such data sets. For instance, if an attribute with a value that sets entity "A" as the organizational attribute, then all users within this attribute are candidate for legitimate use unless further restrictions are introduced. The attribute is of significant importance to protect sensitive content as it will reflect suspicious access/handle of related data sets.

The attribute can be defined by linking the value of the attribute to the complete list of all organizational entities like shown in Table 2.

Attribute value	Organizational entity name
0-0-0	Top management
1-0-0	Drilling
1-0-2	Drilling Engineering
1-0-3	Coring & Fishing
1-1-0	Workover
1-1-1	Workover Engineering
1-1-2	Workover Planning
3-1-0	Exploration Well Appraisal

Table 2: Sample of the Organizational Attributes.

(3) Functional attributes:

These attributes define the main Oil & Gas function related to this dataset in Oil & Gas business. Such information will help to improve the quality of the data and to define security measures and legitimacy of its use.

The functional attribute will position the dataset under “Main Function”, “Sub function”, and “Activity”. For instance, if a given “Well Test” technical report is classified using this taxonomy, the value of the functional attribute for this file will be 5-6-4 as 5 represent the main function: “Reservoir description and dynamics”, 6 represents the sub function: ”Formation evaluation and management“, and 4

represents the activity: “Drillstem or well testing”. Simple examples of main functions and sub-functions are shown in Table 3.

Attribute value	Domain	Function	Sub-function
110	Reservoir Information	Production Surveys	
1202		Seismic and logs	
130		Flow meters readings	
210	Reservoir description and dynamics	Reservoir characterization	
220		Sedimentology	
230		Formation evaluation and management	Drillstem
310	Geologic modeling	Profiles	
320		Processing	
330		Modeling	

Table 3: Sample of the business functions.

(4) Sensitivity Attributes:

This attribute will define the level of confidentiality of the dataset. It will vary from 1 to 4 having 4 as the very high sensitive data and 1 is the non-sensitive data.

The definition of accepted values for each attribute is to be configured by the security framework within the Oil & Gas domain. For instance, the Geographical attribute will be configured to make room for all locations where the business is operating and the sensitivity attribute can be modified to indicate more or less security related sensitivity severity. Sample implementation of the taxonomy for a given file and its attributes is shown in Figure (13).

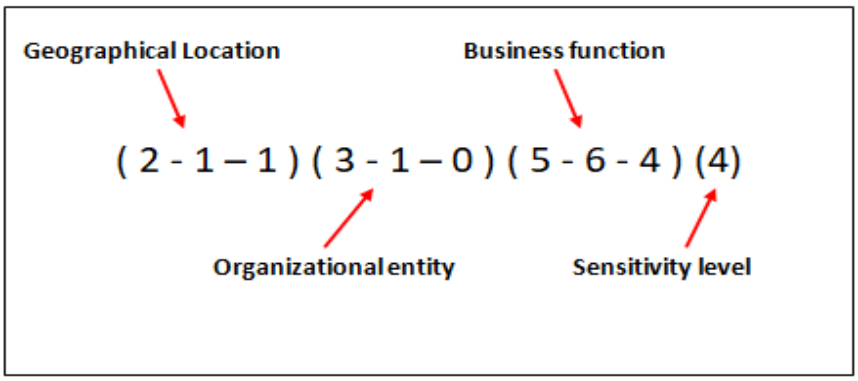

Figure 13: File classified using Oil & Gas information taxonomy.

4.3.3 Text Classifier

Unclassified file (UF) placed in the Unclassified Files List (UFL) will be handled by the Text Classifier (TC) in order to classify the file according to attributes that are defined in the local copy of the Taxonomy (LT). Each Unclassified File (UF) will be given certain value for all attributes. Valid values for each attribute will be fetched from the local copy of the "Taxonomy" (LT) so that classification is up to date. Upon completing the classification task, the Text Classifier will pass the file (UF) along with generated values (LTV1 … LTV4) to the "Classified File List" (CFL) in order for the "File Labeler" (FL) to tag the file. The Text Classifier will

repeat this operation until the "Unclassified File List" is empty. The algorithm shown in Figure 14 describes the operation carried by the "Text Classifier"

File attributes generation

(1) ; ; Gets files from UFL

(2) ; ; Add files to CFL

(3) While UFL ≠ { }

(4) Get file f

(5) *For ∀ LTV_i*

(6) *Generate f_{LTVi}*

(7) *CFL = CFL U f*

(8) *UFL = UFL – f*

Figure 14: File attributes generation algorithm.

Our proposed model will identify the value of each taxonomy attribute for the file in hand. In specific, the classifier will identify the following: (a) Geographical location (b) Organization: top management, drilling, exploration, reservoir management, etc. (c) Function: Reservoir Information, description and dynamics, Geological model, etc., and (d) Sensitivity attribute: The confidentiality level of the content.

Attributed files will be handed over to the File Labeler with a value for each attribute as shown in the figure (15).

The first component in the Text Classifier is the “Content Extractor”. The main function for this component is the extract or converts the content of handled files to text so that it can be handled by the classifier. The component is using programmatic modules according to the file type.

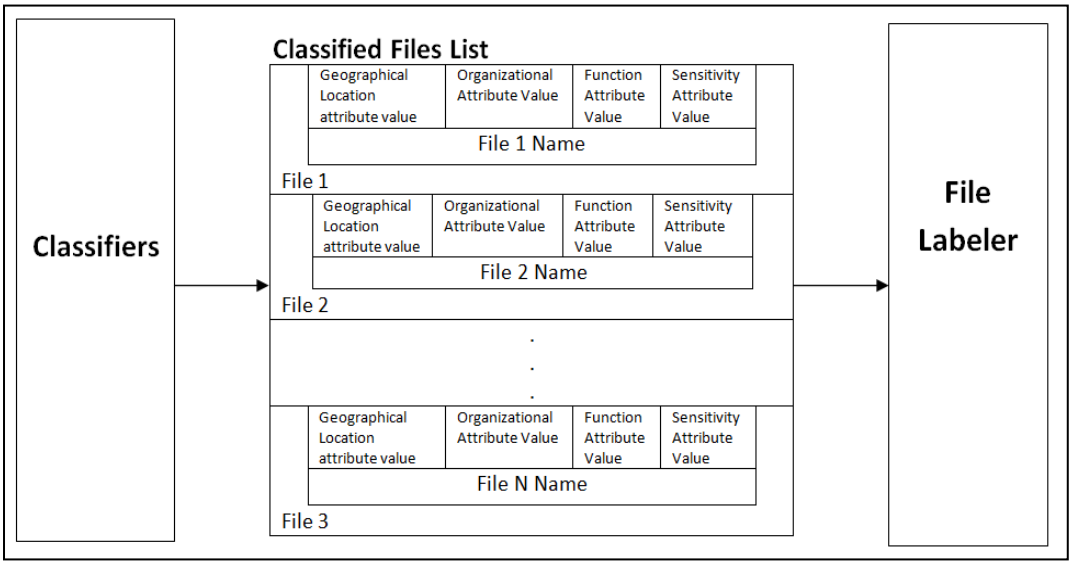

Figure 15: Output of the model’s classifier.

Multiple classification cycles are needed to complete full-classification for each file. As shown in Figure (16), separate classification steps are included in the proposed model for this purpose. The Geographical classifier sub-module will identify the Geographical location; the organization classifier sub-module will identify the file organization and so on. At the end, the file will be sent to the “Classified File List” with the generated value for each taxonomy attribute. Content of file

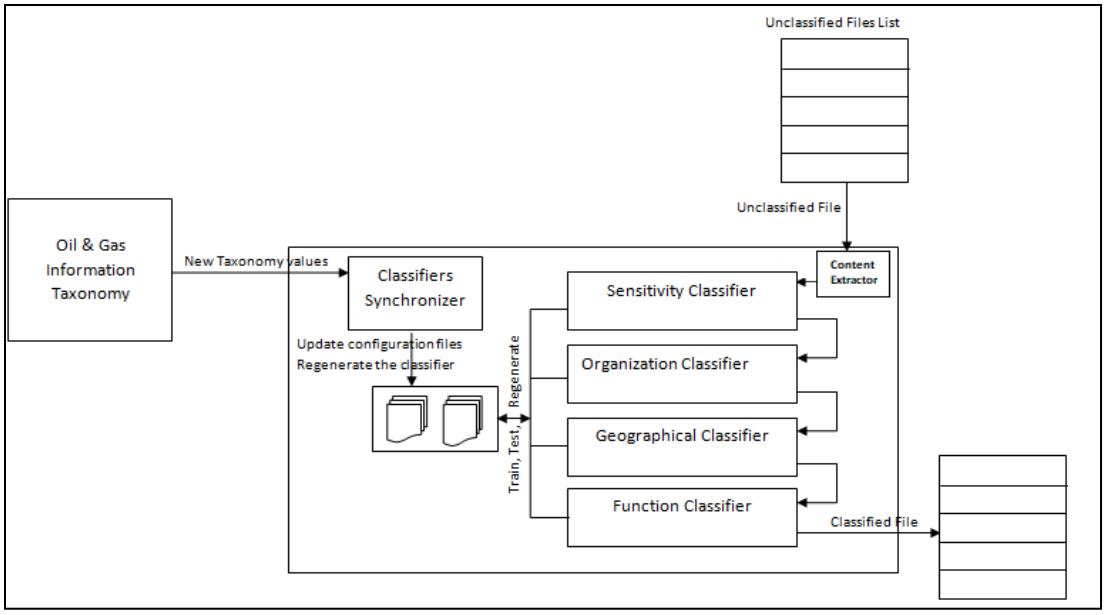

Figure 16: Operations and maintenance of the classifier.

Text Classifiers should be updated all the time to ensure that generated values are reflecting changes on the central taxonomy attributes in a timely manner. This goal is achieved by means of keeping the Text Classifier synchronized with the central taxonomy. Classifier-Synchronizer component will maintain this objective.

The Classifier-Synchronizer CS will compare Timestamp of the Local Copy of the Taxonomy (LT) to check if any changes have been introduced since the last configuration of the Text Classifier (TC). If yes, then CS will perform new configuration cycle on the TC in the following logic:

- Read the new values of the taxonomy.
- Update the classifier configuration files (i.e. features, instances, training set, testing set, etc.)

- Train and Test Classifier sub-module(s).
- Regenerate new version of the TC

If changes are introduced to the Local copy of the Taxonomy (LT), then training and testing datasets might need to be updated as well. New datasets should be supplied by the Global Taxonomy and used by the Taxonomy-Synchronizer to update classifiers configuration files and train them.

The supplied training sets will be used by the classifier to change content into vectors with values of the class that could be any values of the taxonomy attribute. For example, if new geographical location is included, the global taxonomy will be updated and new training set with documents that represent positive and negative examples will be supplied to the synchronizer. The TS will use these documents along with the new taxonomy value to train the relevant classifier(s) so that documents related to that new geographical location will be automatically discovered.

4.3.4 File Labeler

The File Labeler will receive Classified Files (CF) in Classified Files List (CFL). Received files are not tagged (UTF). File Labeler will tag the file with a label that reflects the most recent value for each attribute according to identification done by the Text Classifier. Tagged File (TF) will be written back to its original storage area in (R). Finally, the File Labeler (FL) will update the Classification Timestamp (CT)

and unlock the file. The algorithm shown in Figure 17 describes the operation executed by the File Labeler:

File Labeling and unlocking

(1) ; ; Gets files from CFL

(2) ; ; Add files to R

(3) While CFL ≠ { }

(4) Get file f

(5) *For* $\forall$ LTV_i

(6) *Label f =* LTV_i

(7) Update f_{cts}

(8) *Write f* $\rightarrow$ *R*

(9) *Unlock f*

(10) *CFL = CFL -f*

Figure 17: File labeling steps.

Central taxonomy: ch3anges in the taxonomy attributes and/or values are done on the central taxonomy and synchronized with local copies using push/pull methods. The centralized taxonomy will enable timely changes as required by the business without invalidating the local copies that are used in classifying files in different repositories.

4.3.5 Corporate DLP

The used DLP solution can use a Rule-based method to identify files with sensitive content. Each file that includes “Sensitive Content” Label will be easily identified

by the DLP solution. Appropriate mitigation policy will fire to ensure that unauthorized transaction will not take place.

The fact that all labels are automatically generated with minimum human interaction will ensure that classification is completed in the most accurate, standardized, and consistent manner. All files will undergo same classification according to the defined attributes in the taxonomy.

Files with no labels and/or CT < LCT will be rejected by the DLP solution and send back to the File Checker with Priority = high. This will make the file makes its way to the Unclassified File List (UFL) directly and got classified at the next cycle.

Files with CT > LCT and sensitivity level that matches one of the defined policies in DLP solution will adhere to defined policies.

CHAPTER 5

MODEL PERFORMANCE EVALUATION

We implemented a prototype as a proof of concept to evaluate the effectiveness of our proposed model. The prototype used SVM text classifier and two datasets for configuration (i.e. Training and Testing) and one dataset for evaluation. Results of the prototype evaluation experiment is collected and compared to the results of a standard DLP system in-use. The differences between our prototype results and the standard DLP system results mark the scale of the improvement as suggested by our proposed model.

In the following pages, we will discuss the prototype implementation and the used datasets. We will also explain the configuration of the text classifier and the conducted experiment. We will also show the results of the experiment and the changes in false-positive ration in both scenarios.

5.1 Evaluation Methodology

Since our proposed model has the standard DLP and additional enhancement components, we evaluated our proposal in two different yet integrated areas: (a) the accuracy of the classifier (b) The false-positive ratio of the model. As shown in Figure18, we used two datasets to perform the evaluation. Dataset (1) is used to

evaluate the performance of the classifier and dataset (2) to evaluate the false-positives ratio of the model. Both datasets are explained in details in section 5.2.

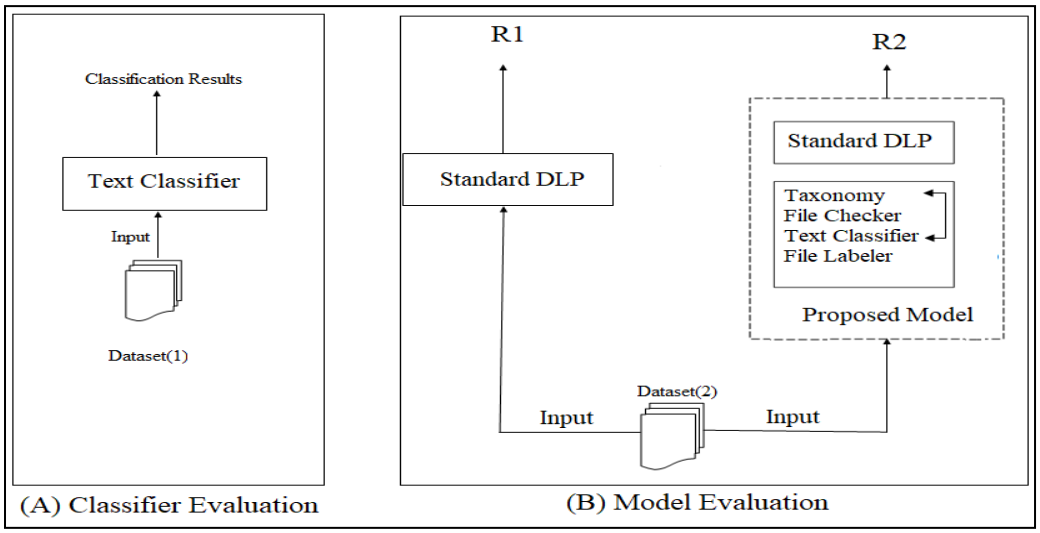

Figure 18: Classifier & Model Evaluation.

The classifier is the major component in the proposed model. It should be configured in a way that provides high accuracy. In order to ensure this requirement, we evaluated the performance of the classifier using testing data within dataset (1). The defined the accuracy as the percentage of the correct classification in these testing files. The evaluation and its results is explained in detail in section 5.3

The model integrates other components with the text classifier. Namely: (a) Information Taxonomy: provides the knowledgebase to be used by the text classifier. (b) File Checker: to identify files that are due for classification, (c) File Labeler: adds metadata to classified files in order to be used by the standard DLP system.

We evaluated the performance of the model using dataset (2). This dataset is used as input to the proposed model and result (R2) is recorded. We compared the correct percentage in (R2) to the percentage of correctness in (R1). The evaluation and its results are explained in section 5.4.

Datasets that are used in the evaluation were carefully selected to give good representation of the population. We performed checks on data files types and sizes to measure the similarity of the datasets with files used in Oil & Gas industry. We provided information about datasets and their content in section 5.2.

5.2 Used Data Sets

We used two datasets in our research as described below:

1- Dataset (1): Total of 2000 files from Oil & Gas domain that we used to train and test our classifiers.
2- Dataset (2): Total of 483 files that were identified as sensitive files by DLP (Rule-based) system in use.

Below is more description about each dataset:

Dataset (1): We collected two thousand documents in English. The collected files represent different files types that are commonly used in Oil & Gas industry. Selected files have different sizes and different formats. Sizes of files vary from 1kb to several MBs. Table 4 shows the percentage of used file formats in our corpora.

Files formats in the corpora included mostly files in PDF format in addition to other common files formats like MS word, MS PowerPoint and MS outlook messages files. These specified file formats represent the vast majority of reports representation in the domain. We checked three file repositories in different sub-domains within Oil & Gas organizations and found that file types in corpora represent more than 95% of all files types in the three repositories.

	Repository 1	**Repository 2**	**Repository 3**	**Total**	**%**
Represented in the selected files types	281,344	124,992	624,321	1,030,657	95.4%
Not represented in the selected files types	6,946	13,701	26,584	47,231	4.6%

Table 4: File types representation in the used corpora.

This finding shown in Figure 19 reflects that selected files in the corpora are a good representation for files used in Oil & Gas domain. Therefore, it will provide good sample for the classifier configuration performance evaluation.

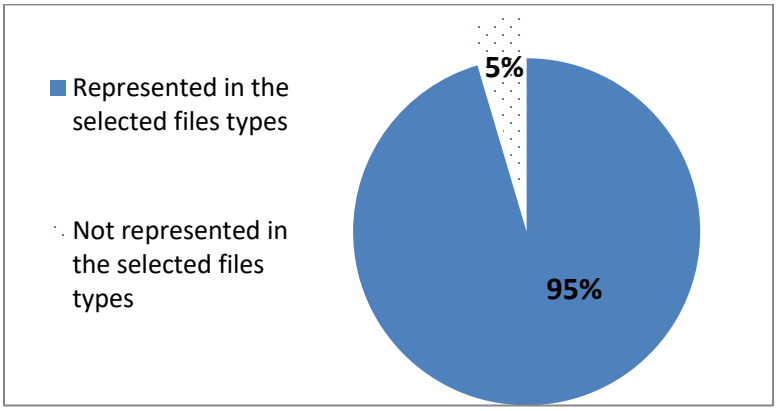

Figure 19: Representation of files in the corpora.

Most of the files in the corpora are PDF files since it is the main format used in generating and distributing final reports. Figure 20 shows that PDF files represent the highest number of selected files. Work in progress files are drafted in MS word files and converted to PDF after been approved and signed by the function owner.

#	File Type	Number of copies
1	PDF files	1000
2	Outlook messages	500
3	MS PowerPoint	300
4	MS Word files	200
	Total	**2000**

Table 5: File formats in the used corpora.

Selected files did not include encrypted or image files as these files will not be covered by our proposed model. Image files can be handled by means of generating a signature for the file. However, this technique has some drawbacks as minor changes in the file will change the signature value and makes DLP system unable to capture it

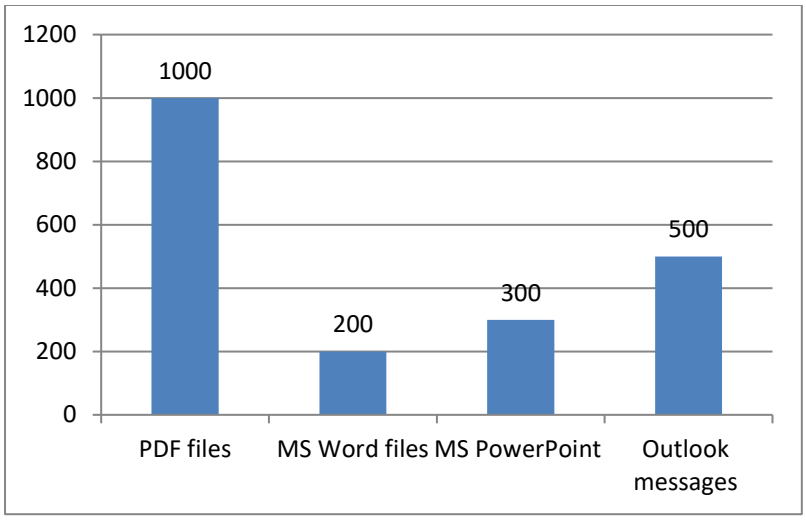

Figure 20: File formats in the used corpora.

Files sizes are in general similar to used files sizes. More than 90% of files (1828 files) in the corpora are less than 2 MB while the rest (172 files) are relatively large files. Figure 21 shows that common files sizes represent the majority in the corpora.

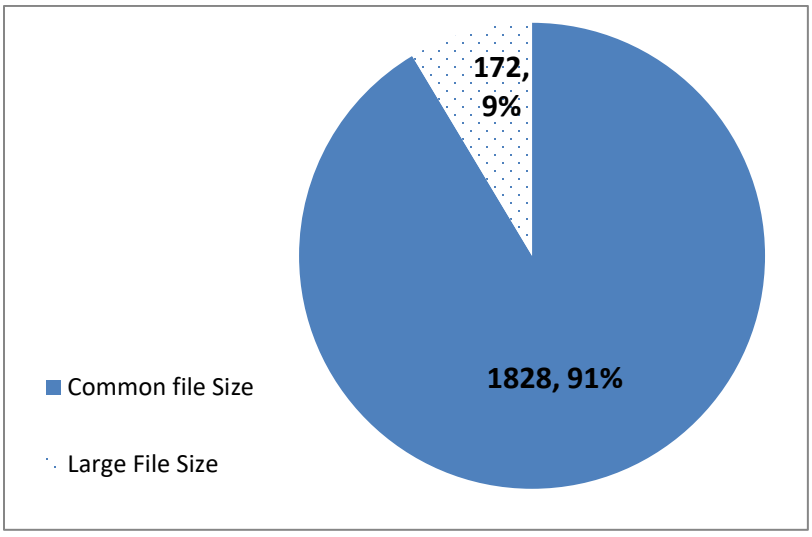

Figure 21: Sizes of files in the dataset.

The files in our corpora are organized into different sensitivity levels according to their content. We asked subject matter experts in the domain to look at the files and

follow specific features and criteria in order to specify the sensitivity level of the content in each file. As a result of this task, all 2000 files were classified and verified by other subject matter expert to make sure that classification is done correctly and consistently.

We verified the consistency of the classification by distributing different copies of the same report over multiple participants and compare their feedback. Inconsistency was limited to one report type and was resolved by updating the criteria used for classification. Final results were consistent among all sub-groups that have been handled by participating subject matter experts.

Table 6 shows the distribution of files according to their content sensitivity. Sensitivity level varies from 1 for non sensitive files to 4 as the highest sensitivity level.

Sensitivity Level	No of Files
Level 1	615
Level 2	824
Level 3	388
Level 4	173

Table 6: sensitivity of files in the data set.

Most of the files are found to be containing sensitive information (i.e. Sensitivity level 2, 3, and 4). Figure 22 shows that files with sensitive content are representing the majority of the files in our corpora

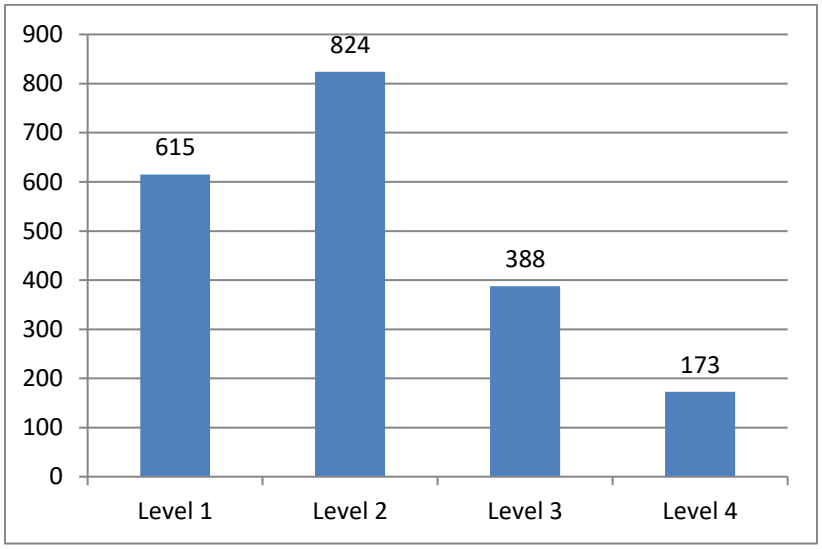

Figure 22: Sensitivity levels of files in the corpora.

The intention of DLP systems is to protect against unauthorized distribution of files with sensitive content outside the corporate network. Therefore, our automated classifier should work to identify files as confidential or not, further classification of files can be completed using different techniques.

According to this generalized classification, all files in the dataset can be described as confidential files or non confidential files. With such classification, confidential files will represent approximately 69% of the total number of files in the corpora as shown in Table 7.

Sensitivity Level	No of Files
Files with Sensitive Content (Confidential Files)	1385
Files without Sensitive Content (Non Confidential Files)	615

Table 7: Representation of confidential files in the dataset.

This situation resembles the situation in Oil & Gas industry where most of the content that is handled by end users is sensitive. Therefore, we believe that collected dataset will give a good representation of the real-life situation and would contribute to the success validation. Figure 23 shows a comparison between confidential and non-confidential files in the corpora.

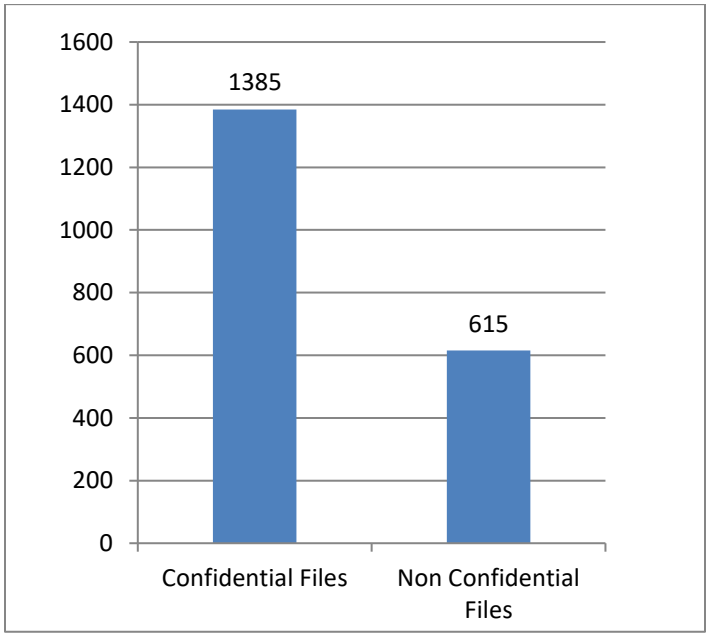

Figure 23: Representation of confidential files in the corpora.

Dataset(2): We collected information about files that are reported by a rule-based DLP solution in Oil & Gas company. Total number of 483 files are identified as

files with a sensitive content. We reviewed the classification specified by the DLP system with subject matter experts. As a result, classification was found to be correct in 67 files only. The rest of files were identified inaccurately and do not contain sensitive content. In other words, we found that false-positive ratio in files specified by the used DLP systems was 86%. Table 8 summaries confidentiality levels of files and false-positive ratio in dataset (2).

Total number of Files	483
Confidential Content using DLP system in use	483
Confidential Content as per further analysis	67
False Positives	416
False Positives percentage	86%

Table 8: False Positives using standard DLP.

We will use this dataset in the evaluation of the model in order to measure the improvement after implementing our proposed model

5.3 Data Pre-processing stage

Data processing includes identifying and preparing your data for the use of the classifier. We intend to use Weka in our text classification. Therefore, we will use Weka configuration file that is called the “arrf” file. This file type is used to configure selected features and classes so that the tool can recognize them. Arrf file has a predefined structure that consists of three sections. Each section starts with @

sign and the section name keyword followed by the configuration value. The first section is called the “relation” section. The section starts by @relation <relation-name>. It is usually used to define and name the relation. The second section is the “attribute” section. This section starts by @attribute <feature-name> followed by the values that can be assigned to this attribute. We also use this section to define our class into either negative or positive values. Generally there are two data types for attributes in arrf files: (1) Nominal data types (2) Numeric data types. The last section in arrf configuration file is the “data” section. It also starts with @data and followed by instances that the classifier will use to learn the classification task. Figure 24 shows the preparation of arrf file and its different section.

Figure 24: Preparation of arrf classifier configuration file.

We completed the following tasks to collect data from the research domain and prepare it for the classifier:

1- Identify files to be used in the experiment:

   We collected 2000 files from Oil & Gas business. Collected files are selected from files repositories within the following sub-domains:

   - Reservoir Management (Production)
   - Exploration
   - Drilling

Selected files are shared with Subject Matter Experts (SMEs) to assess their confidentiality level and label them as “Confidential” or “Public”. The process resulted in identifying and labeling 1385 files as “Sensitive” files and “615” files as “Not Sensitive” files.

2- Identify the instances:

For each file type, we selected some contents that describe the file. Selected content is given the same label as the containing file and considered as the instances for training and testing datasets. In some cases, we loaded the whole file content as an instance.

We started by having 200 instances and continued to add more instances to improve the representation of the data in the domain and/or improve the accuracy of the classification model.

3- Configure the classification tool:

We used Weka as a tool to develop and evaluate the text classifier. This tool uses configuration file of arff extension to specify classification attributes and instance data. We built the file with attributes definition, class definition, and data section. In our configuration, we defined the following attributes: (a) Document of type: String, and (b) class of type nominal (i.e. Values = Confidential, Public).

4- Generate the word vector

We used “StringToWordVector” utility in Weka to generate the matrix of features using the defined train file. The tool generated a vector matrix which

included all distinct words in the data set and allowed us to specify the attribute to be used as the "Class". We performed the following steps to complete this task:

- Loaded the arff file into Weka
- Called StringToWordVector filter and identify the attributes and the class. We configured the features selection on the following way:
  - StopwordsHandler: Rainbow
  - Stemmer: IteratedLovinsStemmer
  - Tokenizer: 1 (Unigram)

5- Generate the classifier

We generated the SVM classifier by selecting SVM algorithm using "linsvm" from within Weka and applied it on the arff file using the training dataset.

6- Classifier accuracy:

We selected content or complete files for testing purposes. Selected content was used to generate 600 unlabeled instances in the arff test file. Then we used the saved model to evaluate the performance.

When we load the test file, we avoided specifying the vector matrix in the test file. Otherwise it will not work. Because the number of vectors in the train and test file must be equal in names and size. The workaround is to load the train file without generating the vector matrix. Instead, we selected the

classification and the filter algorithms in one step. The algorithm has the ability to handle data that is passed through arbitrary filter so that number of vectors will not be considered.

In “Classify tab”, we chose the “Classifier” and set the filter to “StringToWordVector”. Having the tool set this way, it will apply it to Train and Test set together, but if we do it separately, then may be un equal attribute will result in compatibility issue in Weka.

Another workaround is to load all data in one vector matrix and perform the training and testing on that matrix. This way, we ensure that both datasets are of the same number of vectors.

We evaluated the classifier accuracy by noticing the “Classifier output” window and observed the accuracy results. If accuracy result does not satisfy, we repeated steps 2 to 6

In libsvm package, second output parameter of svmpredict() function should give the percentage of performance.

This way, Weka will build a model using the specified Training set and test it over test dataset

When we apply that we will find that we have all as zeros. To know the prediction of the model on the test data, we need to activate an option of “Output Prediction” in Weka and run the model again. In the “Classifier output” window the predictions made by the classifier. For every instance, the classifier will predict the value of “?” as either Confidential or Public.

We measure the correctness by comparing results to our test file to see what predictions are correct and what predictions are wrong.

Second output parameter of svmpredict() function should give the percentage of performance.

7- Save the model.

We saved the model for future use

Loading text files content into arff file”

We used Weka utilty called “TextDirectoryLoader” to load Confidential and Public instances into Weka configuration file (i.e. arff file). This utility will perform the load from positive and negative examples in a complete automatically fashon. Files should be prepared in the repository in a certain way so that the utility can figure our positive and negative examples.

At the end of the automatic load, Weka configuration file will be created and will include instances from all specified files.

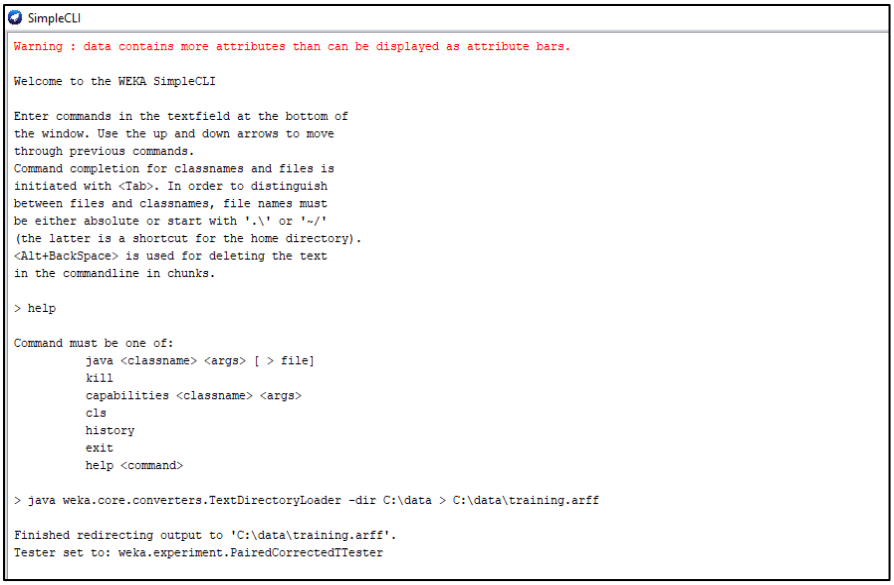

Figure 25: Loading instances from files.

5.4 The classifier Evaluation

We used Support Vector Machine (SVM) with a linear kernel using Weka toolkit to classify files according to their sensitivity level. We used unigram with elimination of stop words and limited the total number of feature to 10,000 in order to control the dimension of our corpora. If more than 10,000 unique non-stop words are found, then we remove noisy/rare features and select the most frequent 10,000 words. We used this configuration as our baseline classifier in our experiment.

We divided files in dataset (1) into training and test datasets. The training data set included 500 files while test dataset included the rest (i.e. 1500 files). In our prototype we limited the number of nominal values (i.e. classes) to four only. We also limited the number of instances to 750. This limitation was imposed for

simplicity reasons in the prototype. More classes can be added in the full implementation of the model and more instances can be included in the training or testing datasets.

We did not include other taxonomy attributes in our prototype. The configuration and implementation of these attributes is identical to the "Sensitivity Classifier" text classifier. The inclusion of all attributes will allow for more automation as interdependency will improve the ability to identify non-legitimate data interchange.

The accuracy of the classifier is percent of the correct classifications. The error rate is the percent of incorrect answers. Thus, accuracy = 1- Error rate.

5.4.1 Features Selection

We loaded all features from instances into the vector matrix. Initially, the number of features was 1677. We eliminated numeric and some special characters so that the number of features became 1511.

Top 10 attributes were:

%	Attribute
0.354633	359 Well
0.343191	924 Cyber

0.313643	1043 Session
0.311612	1056 Symposium
0.300512	954 Friday
0.300512	1270 happens
0.300512	1467 threat
0.300512	1180 cyber
0.291873	641 per
0.291849	471 data

Table 9: Top attributes for feature selection

Then we tried to evaluate the attributes using "CorrelationAttributeEva" with Search function: "Ranker:"

Top 10 attributes were:

%	**Attribute**
0.2981	402 application
0.2981	358 Course
0.2981	368 Engineering

0.2981	419 guidelines
0.2105	206 data
0.2095	416 field
0.2095	413 distributing
0.2095	410 death
0.2095	412 displayed
0.2095	407 computer-related

Table 10: Attributes ranking using CorrelationAttributeEva.

"StringToWordVector" utility & Configuration Parameters

StringToWordVector is a filter utility in Weka that can be used to convert string attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings. The set of words (attributes) is determined by the first batch filtered (typically training data). This filter is not strictly unsupervised when a class attribute is set because it creates a separate dictionary for each class and then merges them.

StringToWordVector can be configured using number of built in Parameters such as Stemmer, Tokenizer, StopwordsHandler, IDFTransform, and LowerCaseTokens. The complete list of attributes along with their usage is explained on Weka documentation pages within University of Waikato website (https://www.cs.waikato.ac.nz/ml/weka/documentation.html)

Stopwords

We used Stopwords list based on Rainbow (http://www.cs.cmu.edu/~mccallum/bow/rainbow/) as shown in the list below:

StopWord	StopWord	StopWord	StopWord	StopWord	StopWord
a	currently	immediate	once	th	who
able	d	in	one	than	whoever
about	definitely	inasmuch	ones	thank	whole
above	described	inc	only	thanks	whom
according	despite	indeed	onto	thanx	whose
accordingly	did	indicate	or	that	why
across	different	indicated	other	thats	will
actually	do	indicates	others	the	willing
after	does	inner	otherwise	their	wish
afterwards	doing	insofar	ought	theirs	with
again	done	instead	our	them	within

against	down	into	ours	themselves	without
all	downwards	inward	ourselves	then	wonder
allow	during	is	out	thence	would
allows	e	it	outside	there	would
almost	each	its	over	thereafter	x
alone	edu	itself	overall	thereby	y
along	eg	j	own	therefore	yes
already	eight	just	p	therein	yet
also	either	k	particular	theres	you
although	else	keep	particularly	thereupon	your
always	elsewhere	keeps	per	these	yours
am	enough	kept	perhaps	they	yourself
among	entirely	know	placed	think	yourselves
amongst	especially	knows	please	third	z
an	et	known	plus	this	zero
and	etc	l	possible	thorough	
another	even	last	presumably	thoroughly	
any	ever	lately	probably	those	
anybody	every	later	provides	though	
anyhow	everybody	latter	q	three	
anyone	everyone	latterly	que	through	
anything	everything	least	quite	throughout	

anyway	everywhere	less	qv	thru	
anyways	ex	lest	r	thus	
anywhere	exactly	let	rather	to	
apart	example	like	rd	together	
appear	except	liked	re	too	
appreciate	f	likely	really	took	
appropriate	far	little	reasonably	toward	
are	few	ll	regarding	towards	
around	fifth	look	regardless	tried	
as	first	looking	regards	tries	
aside	five	looks	relatively	truly	
ask	followed	ltd	respectively	try	
asking	following	m	right	trying	
associated	follows	mainly	s	twice	
at	for	many	said	two	
available	former	may	same	u	
away	formerly	maybe	saw	un	
awfully	forth	me	say	under	
b	four	mean	saying	unfortunately	
be	from	meanwhile	says	unless	

became	further	merely	second	unlikely	
because	furthermore	might	secondly	until	
become	g	more	see	unto	
becomes	get	moreover	seeing	up	
becoming	gets	most	seem	upon	
been	getting	mostly	seemed	us	
before	given	much	seeming	use	
beforehand	gives	must	seems	used	
behind	go	my	seen	useful	
being	goes	myself	self	uses	
believe	going	n	selves	using	
below	gone	name	sensible	usually	
beside	got	namely	sent	uucp	
besides	gotten	nd	serious	v	
best	greetings	near	seriously	value	
better	h	nearly	seven	various	
between	had	necessary	several	ve	
beyond	happens	need	shall	very	
both	hardly	needs	she	via	
brief	has	neither	should	viz	
but	have	never	since	vs	
by	having	nevertheless	six	w	

c	he	new	so	want	
came	hello	next	some	wants	
can	help	nine	somebody	was	
cannot	hence	no	somehow	way	
cant	her	nobody	someone	we	
cause	here	non	something	welcome	
causes	hereafter	none	sometime	well	
certain	hereby	noone	sometimes	went	
certainly	herein	nor	somewhat	were	
changes	hereupon	normally	somewhere	what	
clearly	hers	not	soon	whatever	
co	herself	nothing	sorry	when	
com	hi	novel	specified	whence	
come	him	now	specify	whenever	
comes	himself	nowhere	specifying	where	
o	consequently	concerning	still	whereafter	
his	hither	obviously	sub	whereas	
consider	hopefully	of	such	whereby	

considering	how	off	sup	wherein	
contain	howbeit	often	sure	whereupon	
containing	however	oh	t	wherever	
contains	i	ok	take	whether	
If	ie	okay	taken	which	
could	orresponding	old	tell	while	
course	ignored	on	tends	whither	

Table 11: Stopwords in Rainbow.

5.4.2 Evaluation

To complete the evaluation of the classifier, we reviewed the files in our corpora and prepared them for the use of the classifier. Selected files are originated from different sub-domains in Oil & Gas industry. They included positive and negative instances examples. We firstly converted all files and extract text from them. Then we selected portions from each extracted text in a way that describes the file content. Extracted text was considered as entries in the instance list for the arrf file as shown in Figure 24.

We then worked to clean the data, meaning that we removed unrecognized syntactically rejected characters. We also removed stop words and prepared files entries to be used in vector generation as shown in Figure 26.

File Edit View

Confidential Cleaned - Training.arff

Relation: Sensitivity

No.	1: Document String	2: class Nominal
1	wellhead type well cross section plot production engineering plot by wellbore wb of ltrl latest dpn location (utm)	Confidential
2	New Well Location Review	Confidential
3	Well Testing Unit	Public
4	Review SKETCH SHOWING FACILITIES IDENTIFIED BY WELLSITES . P&EFD SHOULD VERIFY ALL OTHER ...	Confidential
5	the employee responsible for receiving, distributing, tracking, reporting support activities through BMC Remed...	Public
6	DAILY OIL AND GAS PRODUCTION	Confidential
7	DRILLING PROGRAM	Confidential
8	While computer-related health problems are not life death issues, they can be a real pain in the neck. If you ca...	Public
9	Production injecting data are displayed on the application on multiple levels. Starting from well level up to fiel...	Public
10	The following the recommended drilling procedure for well to be drilled as an S-Shape vertical pro...	Confidential
11	production injection	Confidential
12	FORMATION TOPS	Confidential
13	production environment	Public
14	DST Final report for well	Confidential
15	Safety regulation requires that contractor to follow defined procedures as per engineering guidelines	Public
16	surf zone bh, wellhead type :power water injection last drift angle dev tbg min-id (wb0) : wellbore status: sidetr...	Confidential

Figure 26: Loading cleaned data in arrf classifier configuration file into Weka.

We defined the class that will be used for classification upon loading the file into the used tool (i.e. Weka). Identifying the class will enable the classifier to measure the distribution, occurrence and weight for every feature. Figure 25 shows the arrf file after been loaded into Weka with clear distinction between the attribute and the classification class. After that, we generated vectors from selected instances. Total number of features was initially 7061 as shown in Figure 26. Later on, the number of features increased as a result of adding more instances to the arrf file. Adding more instances was needed to correct the classifier results during testing stage.

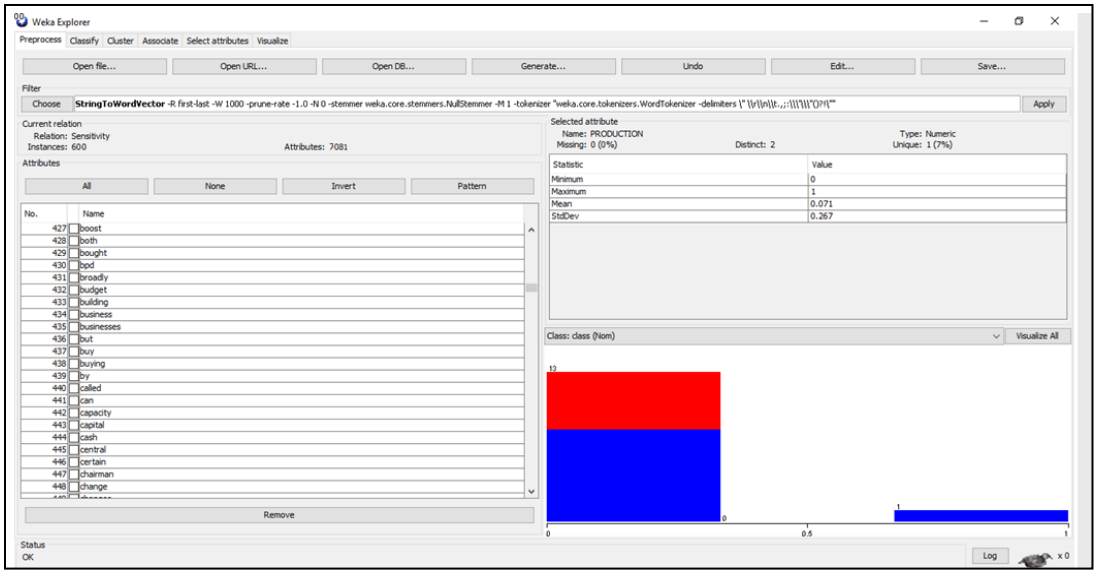

Figure 27: Generating Vectors for SVM classifier.

The classifier was tested in test data of 1500 files. Initial results did not show high accuracy. However, adding more instances to the training list and working with the selected features had resulted in high accuracy rate of 97.1% as shown in Figure 28

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.971     1         0.985      0.5     Confidential
                 0         0          0         0         0          0.5     Public
Weighted Avg.    0.971     0.971      0.943     0.971     0.957      0.5
```

Figure 28 Results of SVM classifier accuracy.

ROC for Algorithm evaluation

ROC analysis was performed after each configuration change in the following:

1- Model settings

2- The instances.

3- Selected features.

Below is the snapshot of performed ROC evaluations:

We picked all attributes to see if the model will perform well. Results was not accepted. Snapshot is below:

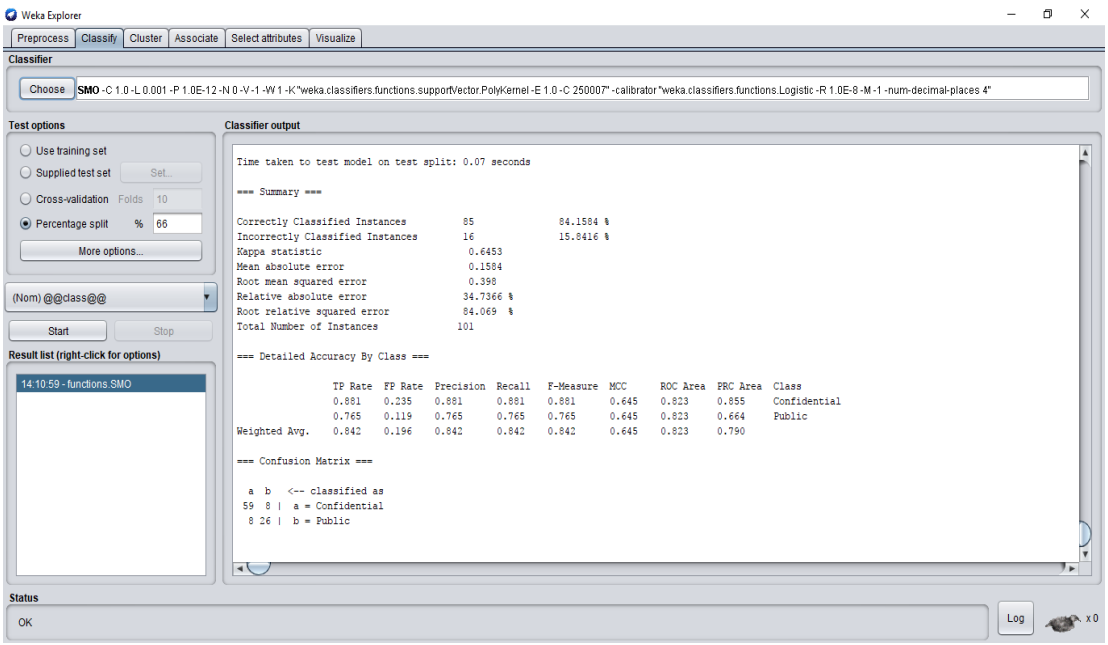

Figure 29: Attributes for ROC analysis.

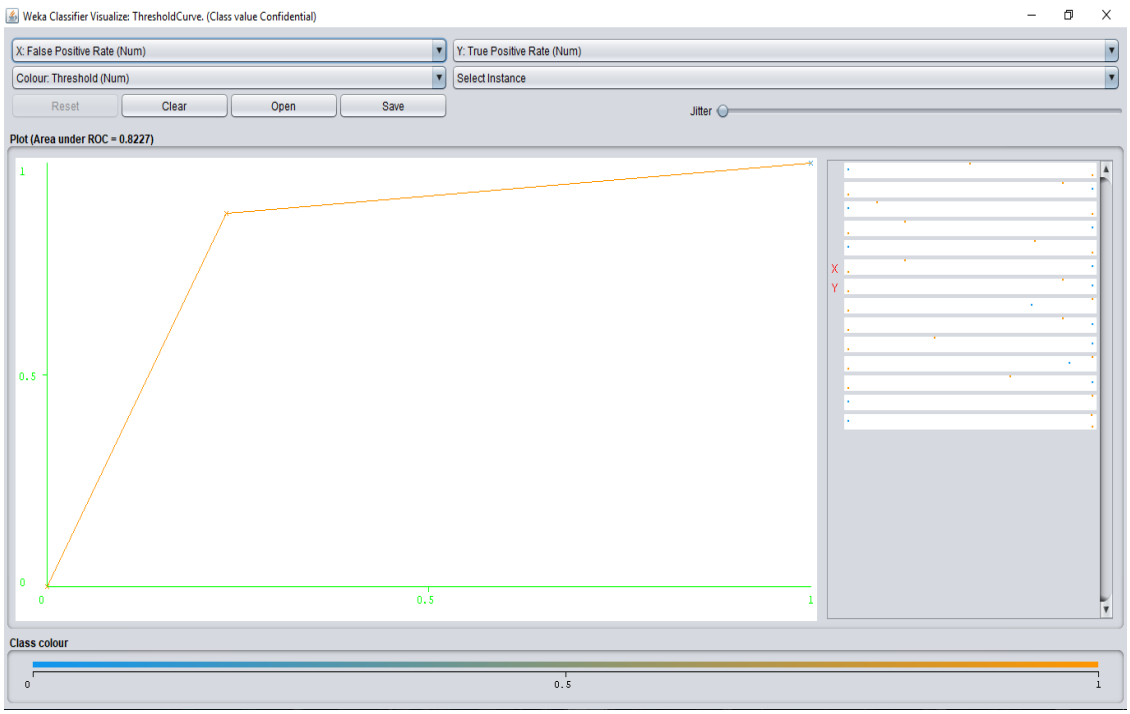

Figure 30: ROC Analysis (1).

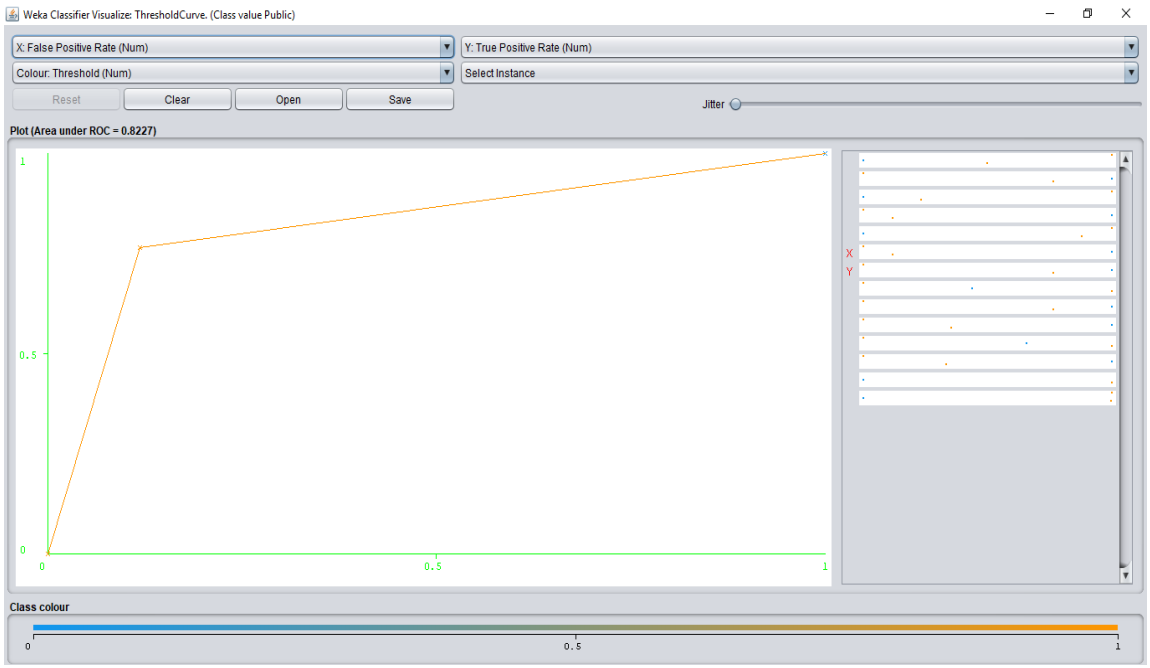

Figure 31: ROC Analysis (2).

Then we removed all numerical values from attributes. Attribute number changed from 1676 to 1511. Same number of instances 296.

Classifier performance was almost the same:

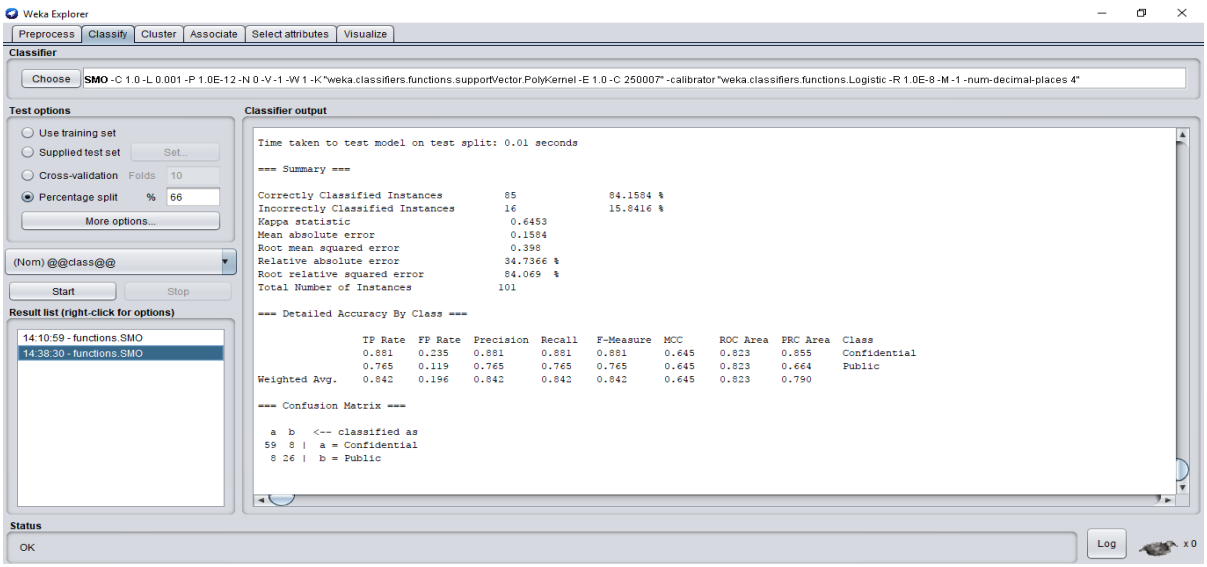

Figure 32: ROC Analysis (3).

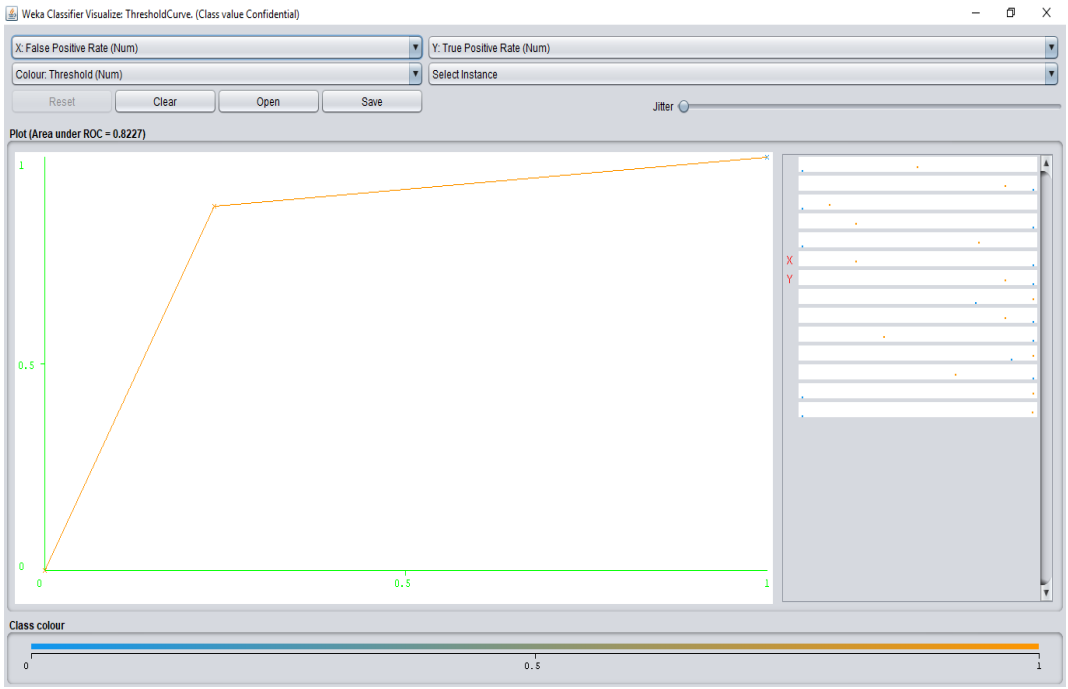

Figure 33: ROC Analysis (4).

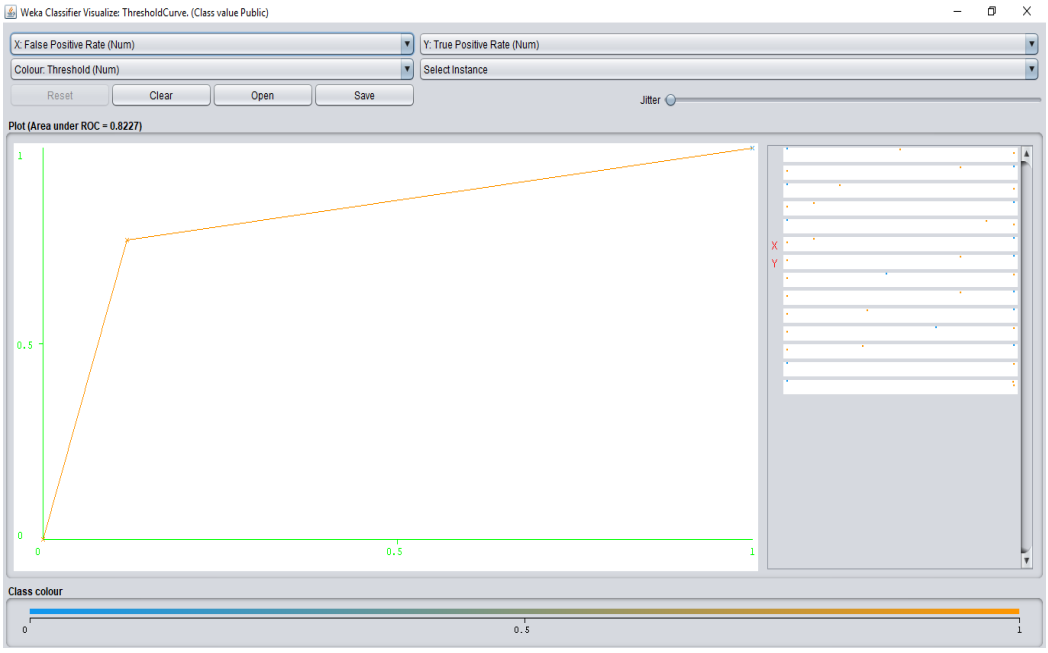

Figure 34: ROC Analysis (4).

Then we change the StringToWordVector attribute: Stemmer to use IteratedLovinsStemmer. This stemmer is an iterated version of the Lovins stemmer. It stems the word (in case it's longer than 2 characters) until it no further changes. Julie Beth Lovins (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics. 11:22-31.

Classifier Performance: No change

Then we changed the “Tokenzer” attribute in StringToWordVector in order to see the impact of handling two or three words together while converting words to vector. Splits a string into an n-gram with min and max grams.

We changed the percent split of training/test instances to be 33/66% respectively.

Results were less accurate:

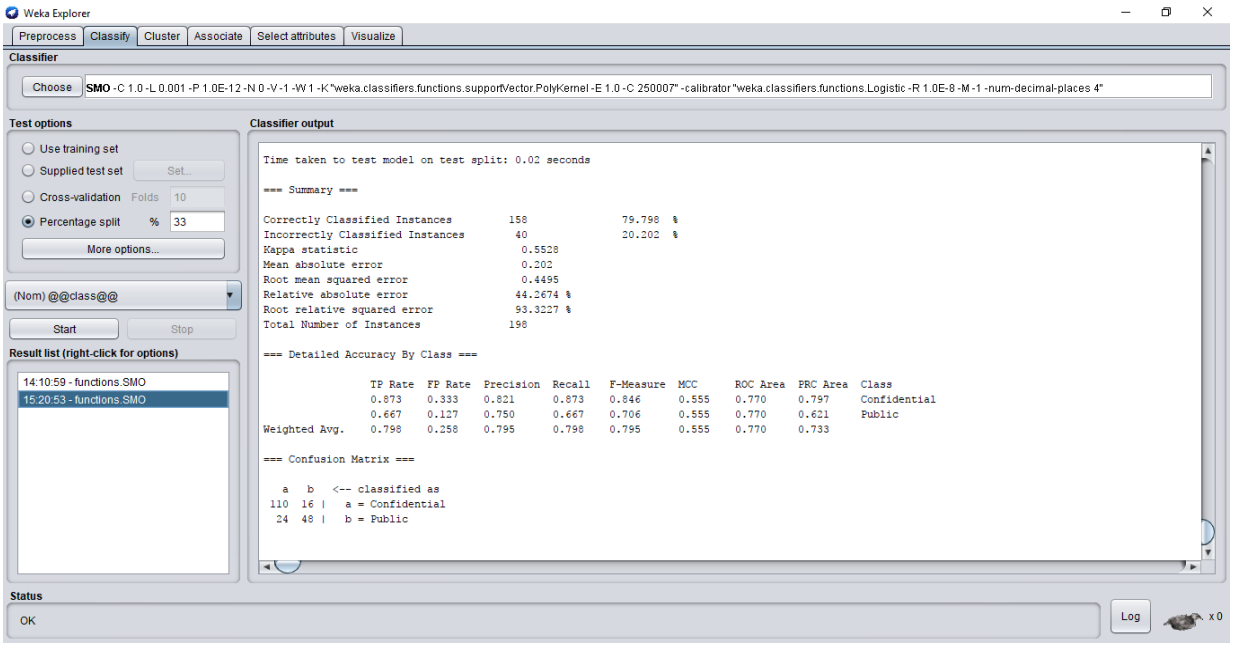

Figure 35: ROC Analysis (5).

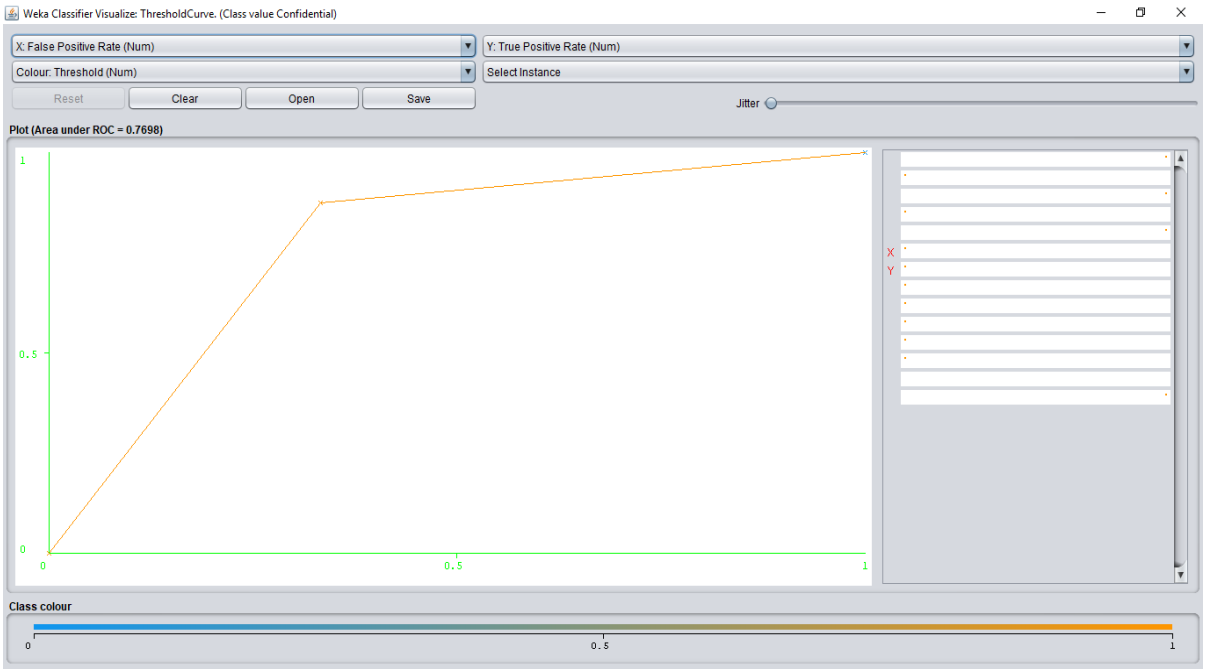

Figure 36: ROC Analysis (6).

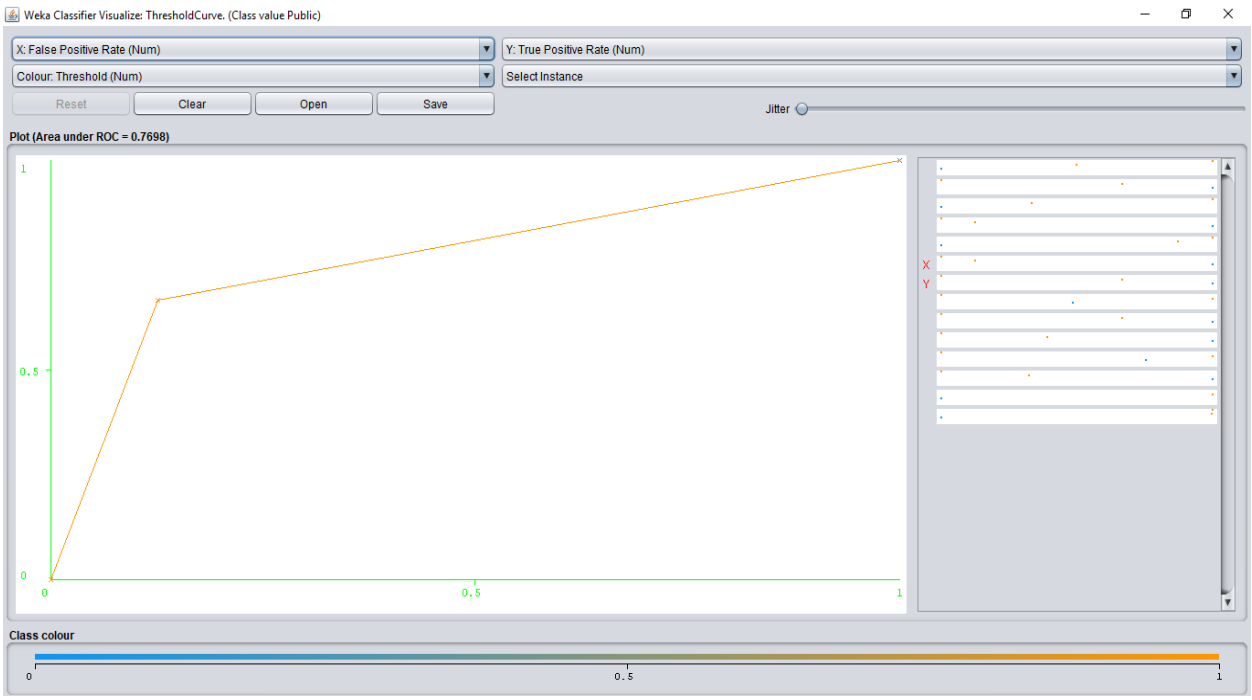

Figure 37: ROC Analysis (7).

With cross-validation: 10 folds instead of percentage split. Results were also less accurate.

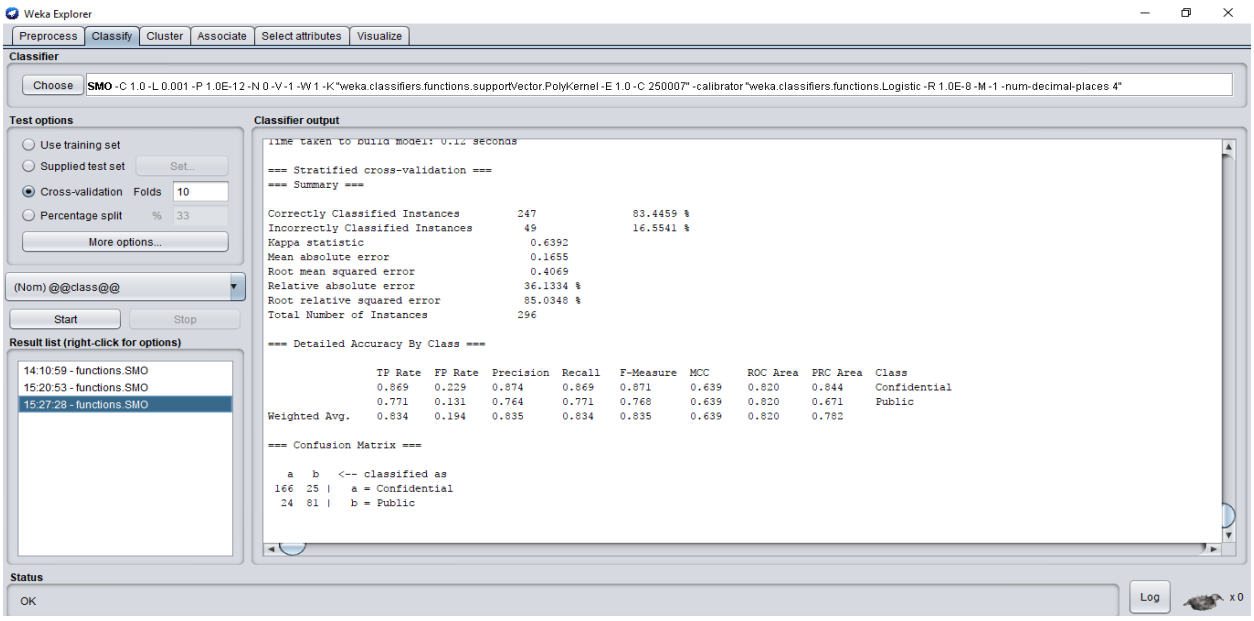

Figure 38: Cross Validation (10 folds).

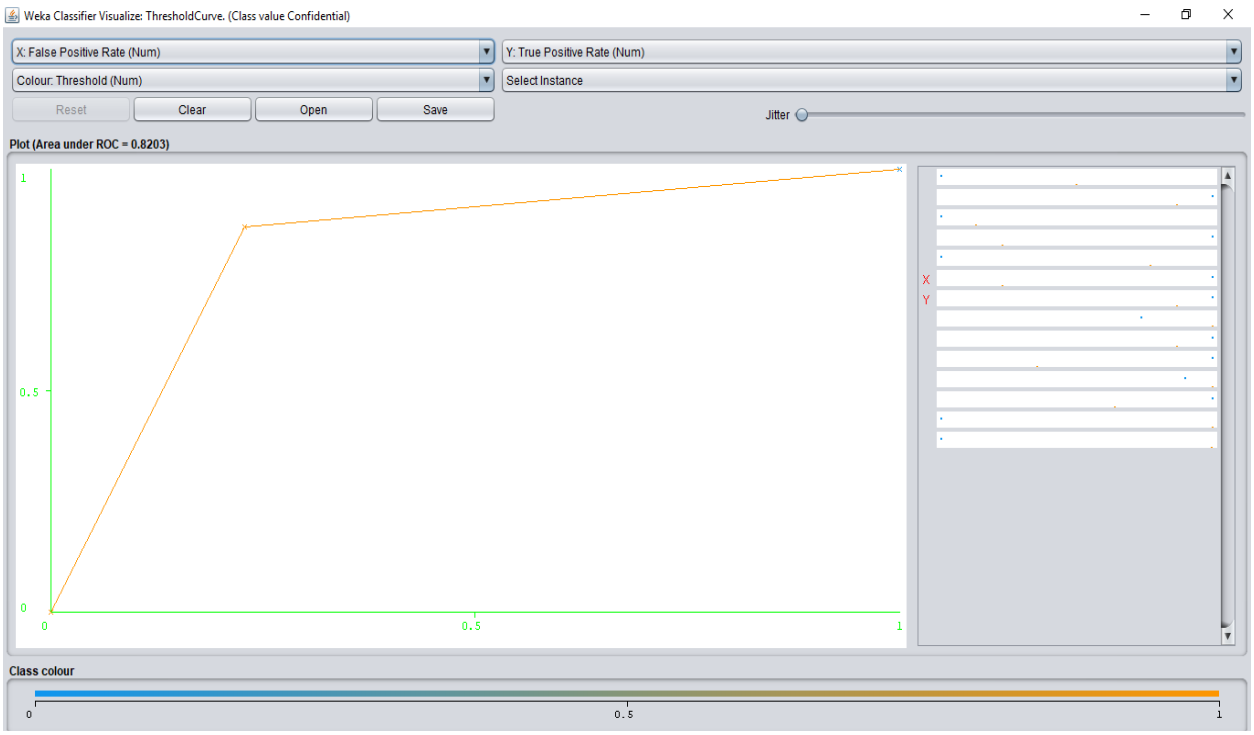

Figure 39: Cross Validation (10 folds) - ROC Analysis.

Model was able to identify 99.6% percent correctly

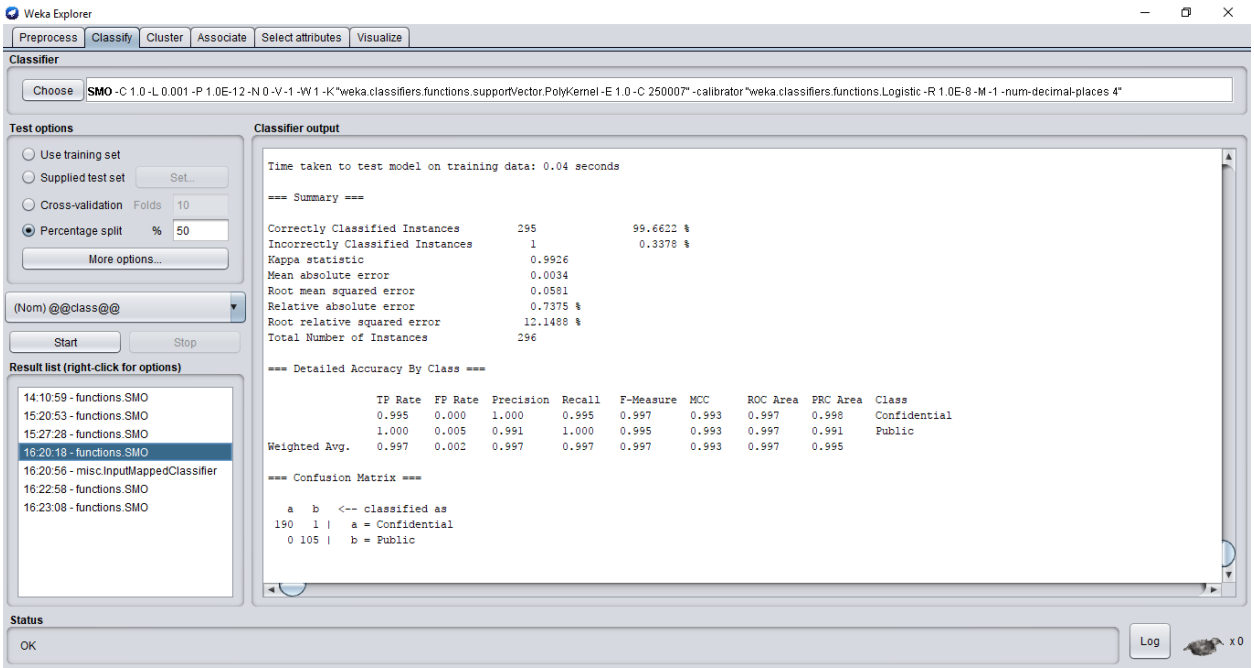

Figure 40: Model Accuracy.

Experiment to evaluate multiple algorithms:

We used Weka to evaluate multiple algorithms. We selected the configuration file (i.e. arff file) and the configured the tool to evaluate NaiveBayes and SVM algorithms.

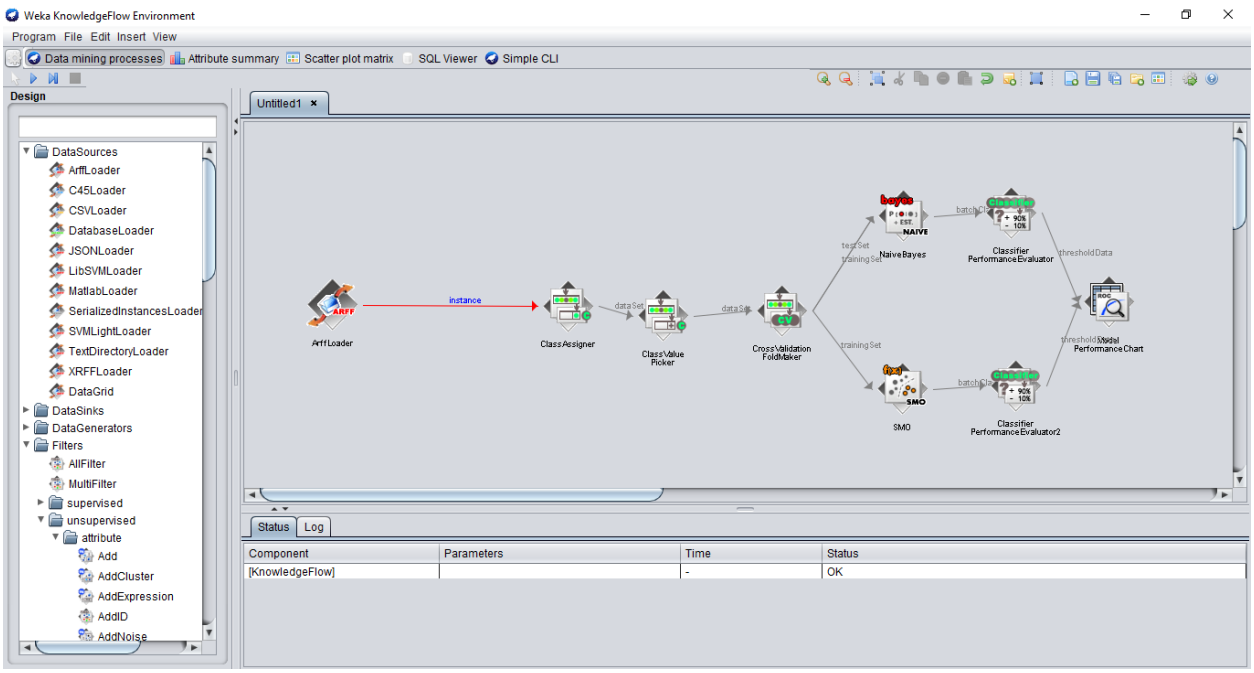

Figure 41: Algorithms Evaluation.

Results of Weka algorithm evaluator showed that SVM algorithm will perform better based on supplied arff file and included instances

5.5 Model Evaluation

Our objective is to validate that the DLP system will perform better using the text classification done by SVM and Domain-specific taxonomy. Thus, we trained and

tested the algorithm using the dataset (1), and then we used it over the Dataset (2) to measure the improvement. Accuracy will be calculated based on results generated from the Dataset (2). We compared false-positives in both scenarios and concluded the results accordingly.

The environment in which we conducted our experiment has a standard DLP system that is used to protect against possible data leakage. The used DLP system is using different techniques to detect possible leakage. The most important method used is the analysis of content sensitivity based on labels that are added to the files by end users using a standard classification/labeling tool. Typically, end-user will perform subjective evaluation of the sensitivity level of the content and label the file using that tool. Later on, other systems that are used for information protection purposes within that environment will refer to this classification in order to decide what actions to take place. If the file that is transferring over the network is labeled with sensitivity level 2, 3, or 4, DLP system will capture the following information: (a) Log of the possible leakage incident (b) Information about the file.

Collected information will be shared with the business organization which the user who initiated the transfer action belongs. That organization then will conduct further investigation on the reported incident and take action if necessary.

We initiated a monitoring window of the standard DLP system findings during a period of four months. During this period, DLP system reported 483 files as possible leakage incidents. All identified incidents were reported to business organizations

for further analysis. Table 9 shows details of reported incidents during the period from September 2016 to December 2016.

Report Month	Reported Incidents	Reported files
Sep 2016	45	141
Oct 2016	25	62
Nov 2016	17	45
Dec 2016	140	235
Total	**227**	**483**

Table 12: files identified by the current DLP as sensitive files.

We collected information about reported files and found that: (1) All files that were reported as leakage incidents have been labeled by end-users as sensitive files. Sensitivity level varied from level 2 to level 4. (2) Files types are similar to the majority of files used in the domain. In specific, 97% of the files are on one of the following: PDF, MS Word, MS PowerPoint, or email messages as shown in Figure 28. (3) All reported incidents included one or more files that are sent via emails to recipients outside the organization.

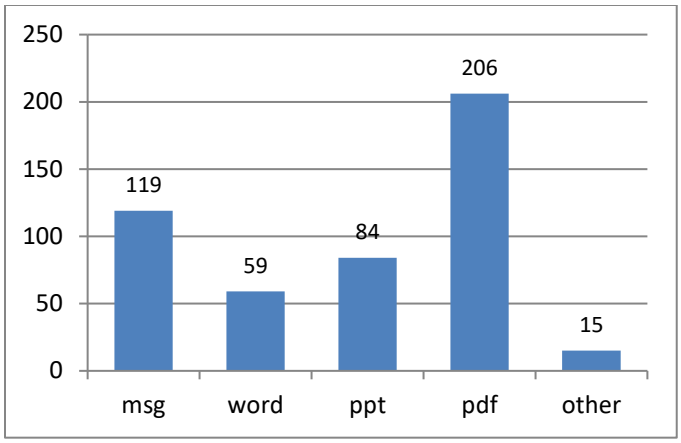

Figure 42: Types of identified files.

We reviewed those 483 files that were identified by the standard DLP system as confidential files with subject matter experts. They identified that only 67 files of them do contain sensitive content. Thus, the standard DLP system has inaccurately identified 416 files as sensitive files. In other words, false-positive ratio was approximately 86% as shown in Table 8.

We used the proposed model with the same set of 483 files. The objective is to find out which files will be specified by the proposed model as confidential files. The proposed model identified 76 files in the list as confidential files. Summary of the identified files using the proposed model is shown in Table 10

The 76 files that are classified by the model as confidential files include the exact list of 67 files that were identified by the subject matter experts as sensitive files. The extra 9 files were inaccurately classified by the classifier as confidential files.

Total number of Files	483
Confidential Content using standard DLP system	483
Confidential Content as per further analysis	76
False Positives	9
False Positives percentage	2%

Table 3: False Positives using proposed model.

5.6 Results Discussion

The proposed model was able to identify same set of files that were identified as confidential files by the standard DLP and the subject matter experts. In addition to that, the number of files that are inaccurately classified was so much less in comparison to the standard DLP system in use. In specific, only 9 files were not classified correctly in comparison to 416 files using the standard DLP.

We are assuming that automatic classification and labeling of the files interchanged over the network will directly change the corporate DLP output. Since that DLP system uses labels by end users, then correcting these labels will directly correct the output and give results similar to the results of the classifier. Therefore, we assume that false-positives at DLP output is equal to 2% as well. We were not able to verify this assumption at the used DLP system due to some legal considerations.

False-positive ratio was decreased from 86% to less that 2% by using our automatic classifier. Table 11 shows the comparison between the classifier results and the standard DLP results:

	Standard DLP	Proposed Model
Total number of Files	483	483
Confidential Content	483	76
Confidential Content as per further analysis	67	67
False Positives	416	9

Table 4: Proposed Model results Vs DLP results.

Figure 29 shows the difference in false-positives between the proposed model and the DLP system in use. We notice that two systems are different at classification stage and that different is reflected at the false-positive ratio.

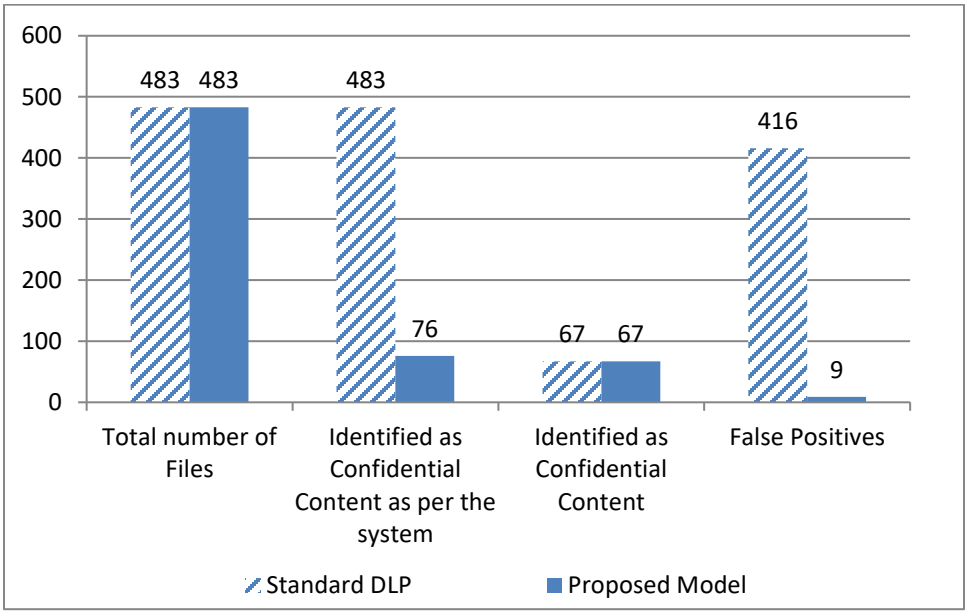

Figure 43: Proposed model results Vs standard DLP results.

Applying the proposed model has resulted in huge reduction in the percentage of false-positives:

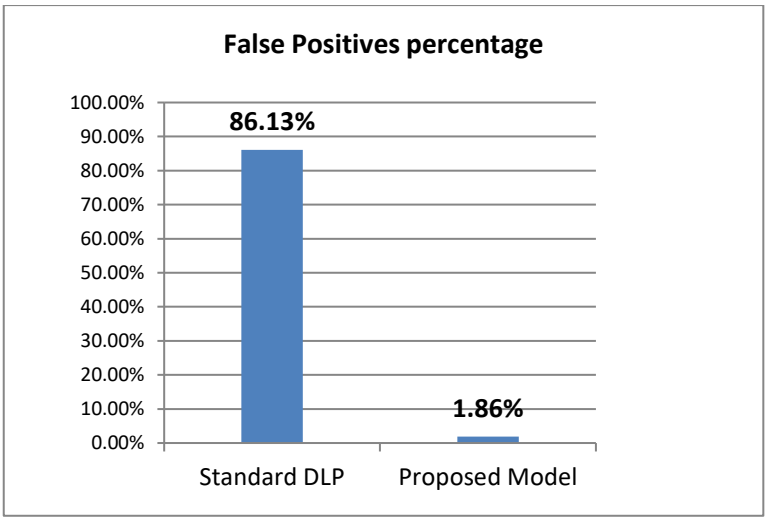

Figure 44: Reduction of false-positive ratio.

The results of our proposed model experiment showed that the use of SVM with domain-specific taxonomy to classify the content will lead to low false-positives rates in classifying files within Oil & Gas domain.

CHAPTER 6

Conclusion and Future Work

Our research aimed mainly to address the problem related to Data Leakage Prevention system within Oil & Gas industry. Namely, we specified the problem of high false-positive ratio. We explained that improving DLP system is a necessity to all business sectors and at same importance to Oil & Gas industry.

We performed a literature review to understand the scale and occurrence of the problem. We also reviewed the literature to identify the suggestions for improving DLP systems by other researchers. We noticed that many researchers have suggested DLP solution can improve by focusing on semantics and context rather than syntax. Some other researchers suggested that DLP solutions can use Labeling methods to have files pre-prepared. We also noticed that rule-based DLPs require less time in configuration and implementation but they suffer from high false-positive ratio in domains where data do not follow predefined syntax. Some researchers suggested that analysis-based DLPs, like text classifiers, will provide an advantage over rule-based DLPs in domains like Oil & Gas industry.

We devised a model that blends together these suggestions and adds to them the use of automatic text classifier with domain-specific taxonomy in a step to eliminate

human factor in the process. Such combination will improve the accuracy of DLP systems and reduce the time to implement them in Oil & Gas organizations.

Our proposed model includes a domain-specific taxonomy that provides comprehensive lexicon of the Oil & Gas industry. We also suggested that understanding the context in which data interchange activity is identified will help the system to decide on the legitimacy of that action. Finally, we proposed that files should be analyzed and labeled prior to the stage of been handled by the corporate DLP.

We suggested that proposed model will improve the performance of DLP systems and is expected to reduce the number of false-positives. It will also improve performance by distributing the processing overhead of subject file over multiple stages. We explained that our proposal includes a domain-specific taxonomy that will enable the standard categorization scheme for classification. Values in the taxonomy will be used by the text classifier to generate meta-data for each file; later on the file labeler will use this metadata to tag files. The applied DLP solution will use updated file tags and file timestamp to decide on appropriate action.

We developed a prototype to evaluate the performance of the proposed model. The experiment we conducted aimed to find false-positives resulted from standard rule-based DLP and compare it with false-positives resulted from our prototype. Our experiment logic is explained below:

- For a sample Oil & Gas (S), use current standard rule-based DLP system to identify false-positive ratio (R1).
- Develop and prototype model (M) that uses: (a) Domain-specific taxonomy (b) Automatic classifier (c) context information.
- Apply the proposed model prototype (M) on (S) to identify false-positives ratio (R2)
- Calculate improvement (I) = R2 - R1.

#	Standard DLP	Proposed Model
1	Relies on manual entry by end users	Automatically classifying documents
2	Centralized processing for captured documents	Distributed processing
3	Extended time for configuration and implementation and	Less time to configure and implement
4	Protect against unintentional data leakage incidents	Protect against unintentional incidents and malicious users
5	Inconsistent and non-standardized	Consistent and standardized

Table 15: Standard DLP Vs Proposed Model.

6.1 Conclusions

We defined a new taxonomy for information within Oil & Gas. The new taxonomy includes information that can be used for different purposes including information security. The new taxonomy comprises of attributes that are specific for Oil & Gas industry and can be configured to fit uses in different organizations in this domain.

The defined taxonomy is referencing business functions as defined by the Society of Petroleum Engineering (SPE) so that it can reflect changes in this domain and continue to be updated all the time. The taxonomy will provide text classification with context related information which will enable more reduction in false-positive ration and meaningful interpretation of identified data interchange activities.

We also defined an improved model for DLP by including a pre-step that prepares the files for an efficient monitoring by DLP components. Based on the new model, the system will bypass the time-consuming discovery part, and the error prone human classification. Instead, it will use a classification based on automatic text classification using SVM and Information taxonomy that covers all business functions in Oil & Gas industry.

Having these components embedded together in one DLP model will improve the reliability of DLP systems in Oil & Gas domain as the model will:

1- Reduce the false-positive ratio.

2- Reduce the time for system configuration.

3- Reduce the dependency on human factor

4- Provide domain-specific knowledgebase and context information

DLP systems are facing three major challenges. These challenges can be summarized as follows:

1- Difficulties in accurately description information to be monitored.

2- Extended implementation time.

3- Limitations in handling encrypted and/or graphical information.

In this research we showed that accuracy in identifying sensitive content in documents within Oil & Gas domain can be resolved by devising a model that uses automated classification algorithm. The algorithm along with Information taxonomy will produce labeled files so that rule-based DLP systems can accurately address files with sensitive content.

In our experiment, we found that false-positive has decreased from 86% to less than 2% which will enable DLP systems to overcome accuracy challenge

The second challenge of extended and complicated implementation of DLP solution can be resolved by overcoming high false-positives associated with the use of rule-based DLPs. Rule-based DLPs can be easily configured since they require simple configuration. Defining keywords in the taxonomy will enable for two tasks, the content classification and the policy-matching. The taxonomy provides the lexicon

of Oil & Gas industry. It should require minimum adjustments in order to match the Oil & Gas organization that is configuring DLP.

In addition to that, distributing the task of content inspection and transferred file interception over multiple stages in the system operation model will balance the processing overhead and adds to the improved performance of DLP systems in general.

Identifying encrypted files and image files remain an open area for researches. Currently, a few methods are valid for these type of files. Only hashing and signature method can be used to monitor such files. The problem of these methods that users can make a minor change in the file will change the hash-function value and in turn, the file becomes unidentifiable by DLP solutions.

6.2 Future work

The problem of identifying non-classifiable documents is still stands as area for future research. Classifier is not able to identify sensitivity of images and files in encrypted format.

Classifying images is one of the most challenging tasks [45]. Currently, there are different ways to handle this challenge using classification, Chaitali Dhaware et al. [46] surveyed the literature and listed some suggested ways to classify files that contains complete this task: Support Vector Machine (SVM), Artificial Neural Network (ANN) and Decision Tree (DT).

Data Leakage Prevention systems should include the classification modules as a genuine component in the system so that system configuration and setup will not consume more time and effort.

Information Taxonomy for Oil & Gas is also open for more researches as the taxonomy presented in this research doesn't clarify interdependencies between different business functions in the domain. Knowing these interdependencies will help specifying paths in which data-flow is expected. Such information will help reduce false-positives and allow for better protection.

Introducing the domain-specific taxonomy is a valid idea for other industries as well. Files in each industry have commonalities that can be collected, analyzed, and organized in a structure that supports automatic file classification tasks. Researchers should look at industries where the majority of interchanged data is confidential and apply the domain-specific taxonomy. Some examples for implementation are health sector, banking sector, military, etc.

Information taxonomies can also be linked directly to the used classifiers in order to tune their work and ensure that output accuracy is high. Elements of the taxonomy can be assigned different weights and have the classifier directly linked to it. In case of feature selection is needed, elements in the taxonomy with low weight will be the first candidates.

References

[1] Blanke. W.J, " Data loss prevention using an ephemeral key", High Performance Computing and Simulation (HPCS), 2011 International Conference on 4-8 Jul 2011

[2] Sabah Al-Fedaghi, "A Conceptual Foundation for Data Loss Prevention," International Journal of Digital Content Technology and Its Applications. Volume 5, November 3, March 2011.

[3] B. Hook, "Data leakage prevention: Reducing risk", SC Magazine, May 2009.

http://www.scmagazineus.com/data-leakage-prevention-reducing-risk/article/136039/#

[4] P. Kanagasingham, "Data loss prevention", SANS Institute, August 15, 2008.

http://www.getadvanced.net/pdfs/SANS%20Institute%20Data_Loss_Prevention.pdf

[5] How Employees are putting your Intellectual Property at risk, White Paper, ponemon Institute

[6] Liu, Simon & Kuhn R, "Data Loss Prevention", IT Professional (Volume:12 , Issue: 2) March-April 2010 – IEEE 1520-9202

[7] Malte Wedel, Andreas Roessler, "Data loss prevention", US patent 7185238 B2

[8] El Kharboutly, R.; Gokhale, S.S, "Data Loss: An Empirical Analysis in Search of Best Practices for Prevention", Cloud Engineering (IC2E), 2014 IEEE International Conference on 11-14 March 2014.

[9] Borders, K., " Quantifying Information Leaks in Outbound Web Traffic", Security and Privacy, 2009 30th IEEE Symposium, May 17 – 20, 2009

[10] Wuchner. T, " Data Loss Prevention Based on Data-Driven Usage Control", Software Reliability Engineering (ISSRE), 2012 IEEE 23rd International Symposium on 27-30 Nov. 2012

[11] Liwei Ren et al. Document fingerprinting with asymmetric selection of anchor points. US patent 835947

[12] Hongju Yeom & Hwasung Kim, "An efficient multicast mechanism for data loss prevention", Advanced Communication Technology, 2005, ICACT 2005. The 7th International Conference on 21-23 Feb. 2005

[13] Hongju Yeom & Hwasung Kim, " Design of internal information leakage detection system considering the privacy violation", Information and Communication Technology Convergence (ICTC), 2010 International Conference on 17-19 Nov. 2010

[14] Hashimoto, G.T. et al., " A Security Framework to Protect against Social Networks Services Threats", Systems and Networks Communications (ICSNC), 2010 Fifth International Conference on 22-27 Aug. 2010

[15] Lawton, G., " New Technology Prevents Data Leakage", Computer (Volume:41 , Issue: 9) Sep. 2008

[16] Gao Zhi-Min & Wang Sheng-Yuan, " Survey of Information Security Risk Assessment", Electrical and Control Engineering (ICECE), 2010 International Conference on 25-27 Jun 2010

[17] Asosheh, A. & Khani, A., "A new quantitative approach for information security risk assessment", Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[18] Doss, G. & Tejay, G., "Developing insider attack detection model: A grounded approach", Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[19] Moskovitch, R. et al., "Identity theft, computers and behavioral biometrics", Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[20] Koppel, M. et al., “Automatically Classifying Documents by Ideological and Organizational Affiliation”, Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[21] Azarkasb, S.O. & Ghidary, S.S., “New approaches for intrusion detection based on logs correlation”, Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[22] Hoss A. M. & Carver D. L., “Weaving ontologies to support digital forensic analysis”, Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[23] Azarkasb S. O. & Ghidary S. S, “New approaches for intrusion detection based on logs correlation”, Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[24] Finin T et al., “Assured Information Sharing Life Cycle”, Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on 8-11 Jun 2009

[25] Zhihu Wang & Haiwen Zeng, “ Study on the risk assessment quantitative method of information security”, Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on 20-22 Aug. 2010

[26] Tianshui, Wu; Gang, Zhao, “A New Security and Privacy Risk Assessment Model for Information System Considering Influence Relation of Risk Elements”, Broadband and Wireless Computing, Communication and Applications (BWCCA), 2014 Ninth International Conference on 8-10 Nov 2014.

[27] Gen-Yih Liao & Chen-Hwa Song, “Design of a computer-aided system for risk assessment on information systems”, Security Technology, 2003. Proceedings. IEEE 37th Annual 2003 International Carnahan Conference on 14-16 Oct 2003

[28] Jeffrey Hunker & Christian W. Probst, “The Risk of Risk Analysis”, WEIS 2009 * June 25, 2009

[29] Carl A. Foerster, “Analysis of Decision Factors for The Application of Information Access Controls Within The Organization”, Ph.D. Dissertation, George Washington university

[30] Preeti Raman et al. Understanding Data Leak Prevention

[31] N. Vachharajani, M. J. Bridges, J. Chang, R. Rangan, G. Ottoni, J. A. Blome, G. A. Reis, M. Vachharajani, and D. I. August, “Rifle: An architectural framework for user-centric information-flow security,” in Proceedings of the 37th annual IEEE/ACM International Symposium on Microarchitecture

[32] S. Lee et al., “Data leak analysis in a corporate environment,” in Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC ‟09) June 2009

[33] J. Han. “Data Mining: Concepts and Techniques”. USA: Morgan Kaufmann Publishers, 2005.

[34] Japan Network Security Association; Fiscal 2007 Information Security Incident Survey Report. 2008 http://www.jnsa.org/en/reports/incident.html

[35] HDE: Survey of Erroneous email transmission in Japan, 2008 http://www.hde.co.jp/reposrts/20080423

[36] Campbell K, Gordon L, Loeb M, Zhou L. The economic cost of publicly announced information security breaches: empirical evidence from the stock market. J Comput Secur 2003;11:431e48.

[37] Robert Layton a, Paul A. Watters, “A methodology for estimating the tangible cost of data breaches”, 2214-2126/Crown Copyright © 2014 Published by Elsevier Ltd

[38] Rohit Pol, Vishwajeet Thakur, Ruturaj Bhise, Prof. Akash Kate / International Journal of Engineering Research and Applications (IJERA)

ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 3, May-Jun 2012, pp. 404-410

[39] Sandip A. Kale C , Prof. S. V. Kulkarni / Journal of Computer Engineering (IOSRJCE) Vol. 1 Issue 6 (Jul-Aug 2012), pp. 32-35

[40] Radwan R. Tahboub, Yousef Saleh / NNGT Journal: International Journal of Information Systems Vol 1 (Jul 2014)

[41] Machine Learning Sets New Standard for Data Loss Prevention: Describe, Fingerprint, Learn [White Paper]. (2010, December 14). Symantec. [Online]. Available: http://eval.symantec.com/mktginfo/enterprise/whitepapers/b-dlp machine learning.WP en-us.pdf

[42] CA DLP: information protection and control [Product Sheet]. (2010). CA Technologies [Online]. Available: http://www.ca.com/~/media/Files/productbriefs/dlp-12-5-ps.pdf

[43] Trend micro data protection: Solutions for privacy, disclosure and encryption [White Paper]. Trend Micro [Online]. Available: http://us.trendmicro.com/.../datalossprevention/wp02_dlp-compliance-solutions_100225us.pdf

[44] McAfee host data loss prevention [Data Sheet]. McAfee [Online]. Available: http://www.mcafee.com/us/resources/data-sheets/ds-host-data-loss-prevention.pdf

[45] Inur Shazwani Kamarudin et al. / comparison of image classification techniques using Caltech 101 dataset / Journal of Theoretical and Applied Information Technology Vol 71 January 2015

[46] Chaitali Dhaware, Mrs. K. H. Wanjale / Survey On Image Classification Methods In Image Processing / Trends and Technology (IJCS T) – Volume 4 Issue 3, May - Jun 2016

[47] Ken Perkins/ Computer and Information Security Handbook

[48] Ernst & Young's Advisory Services , EY Publications

[49] Understanding and Selecting a Data Loss Prevention Solution - SANS Whitepaper, The SANS Institute http://sans.org & Securosis, L.L.C., http://securosis.com

[50] Prathaben Kanagasingham /Data Loss Prevention, SANS Institute , InfoSec Reading Room

[51] The global Information Systems Security report 2015, 2016 EY Publications

[52] Håvard Devold, "Oil and gas production handbook: An introduction to oil and gas production, transport, refining and petrochemical industry" ISBN 978-82-997886-3-2

Vitae

Name	:Ashraf Eltahir Mohamed Ahmed
Nationality	:Sudanese
Date of Birth	:8/27/1975
Email	:ashraf.tahir@yahoo.com
Address	:P.O. Box 1972, Omdurman 2944, Sudan
Academic Background	:B.Sc.(Honor) in computer Science, Future University, Khartoum, Sudan